Psychometrika

CONTENTS

VOLUME ELEVEN SEPTEMBER 1946 NUMBER THREE

PSYCHOMETRIKA—VOL. 11, NO. 3
SEPTEMBER, 1946

AN APPROACH TO THE PROBLEM OF DIFFERENTIAL PREDICTION

HUBERT E. BROGDEN
THE ADJUTANT GENERAL'S OFFICE*

A procedure for maximizing selective efficiency is developed for application to situations in which it is desired to select from a single group of applicants for several possible assignments. The problem of comparable units for the several criteria whose values must be compared to each other for differential assignment purposes is discussed. It is demonstrated that, assuming linear regressions, maximal selection is obtained if individuals in any given assignment are differentiated from those rejected according to critical rejection scores on the multiple weighted sum of the predictors and from another possible assignment by critical difference scores which are merely the differences between the two critical rejection scores. Since the relationships just indicated give no way of determining the magnitude of the critical scores required to select the required number of persons for each assignment, a successive approximation procedure for accomplishing this purpose has been devised and a computational example is worked out.

The procedures for obtaining maximum efficiency in selecting personnel by means of test scores or other predictors are simple and well known when a single assignment is involved. So far as the author is aware, no procedure has been devised for maximizing efficiency of selection and assignment when each individual may be eligible for several assignments. The present paper will be concerned with presentation of such a procedure.

Before attempting to formulate the problem in mathematical terms, the question of comparability of the units for the criteria of the several assignments will be given some consideration. While criteria in standard score form might be regarded as comparable, this solution involves the tacit and undesirable assumption that all criteria are equal in both variability and importance. In certain assignments the nature of the work may be such that all individuals produce very nearly the same amount, while in other assignments considerable variation may occur. It would, of course, be advantageous to place an individual equally good at both types of work in the latter assignment. Similarly, jobs vary in importance to over-all effici-

* The opinions expressed are those of the author and are not to be construed as official or as those of the War Department.

ency of the organization. Thus, if an applicant for employment in a newspaper office were highly but equally skillful at both sweeping floors and operating a linotype machine, it would be desirable to have him operate the linotype machine. Variation in the efficiency with which the linotype machine is operated affects over-all efficiency of the newspaper office much more than variation in the efficiency with which floors are swept. The following discussion of the characteristics of a meaningful criterion is pertinent to the problem of the comparable units and leads rather directly to the solution which seems to the author to be most desirable. While it will be assumed in this discussion of criteria that the classification problem is that of an industrial concern, most of the comments are with some modification pertinent to classification problems of organizations such as the United States Civil Service Commission or the Army.

Although standard scores were considered defective from the viewpoint of comparability, it is definitely desirable to employ mean deviates, and it will be assumed hereafter that all variables are expressed in such terms. The only reasonable alternative to mean deviates would be variables expressed in terms of absolute zero. Apart from the impracticability of attempting to determine absolute zeros in our present stage of development of selection procedures, the information made available by determining absolute zero is not directly pertinent to the problem of making the best possible selection and assignment from the available applicants. In most selection problems the essential comparison is between the given individual score and the expected score if one were to choose at random from the sample of applicants. This expected score is of course the mean. Selection of an individual producing ten units more than the average applicant would effect a saving of exactly ten units no matter whether the mean were twenty above absolute zero or one hundred above it.

Usually in selection problems a test or a battery of tests is employed to identify in a group of applicants those individuals who will perform most efficiently on the job. Presumably if time and expense were not important, the criterion itself would be employed as the selector. Ideally, a criterion would indicate the difference between the cost to the employer per unit of produce or per service rendered by the given individual and the average applicant (that is, it would be expressed in mean deviate form) multiplied by the number of each of the various types of units called for in the job that the given individual could be expected to produce in a given time unit. The criterion should allow for errors, training costs, turnover, or for any additional cost accounting factors which might be related to individual differences in the abilities or traits of the persons on whom the criterion

scores are obtained. For example, certain types of overhead would be reduced with more efficient producers in some instances. Thus the criterion would indicate the total saving (or loss) to be obtained by selecting the given instead of the average applicant. If the criterion were expressed directly in terms of a reliable measure of the saving in dollars, the obtained criterion scale would be in units having the same meaning at all points of the scale. In addition, units of several criteria so expressed would have the same meaning, or we might say, selective significance, and could be directly compared not only with other units on the same criterion scale, but with any unit on any criterion scale. That is, if individual X in Job A could be expected to perform such that a saving of one hundred dollars would be expected (over the average individual) and in Job B such that a saving of two hundred dollars would be effected (the criterion values would be +100 and +200, respectively), it could be said that a saving of one hundred dollars would be obtained by placing the individual in Job B. If, furthermore, successive pairs of individuals were to have criterion values of —100 and —98, 105 and 107, and 95 and 97 in Job A; and —22 and —20, 117 and 119, and 205 and 207 in Job B, the differences in desirability for employment between these successive pairs would be exactly the same as far as the employer is concerned. In that sense, then, it can be said that the criterion units are comparable.

It is realized that the reasoning here is from the employer's viewpoint only, and that there are, even so, many intangibles that would never be expressed in monetary terms. In fact, a close approximation to such a criterion could probably not be obtained. However, consideration of the desirable characteristics of the criterion may help in obtaining the best approximation possible under the given circumstances. The desirability of obtaining such an approximation would be much more evident in a differential prediction problem than in those involving a single criterion, since, in the former, the problem of comparing units of different criteria is added to that of comparing units at different parts of the scale of the same criterion. Insofar as the units and scales are not comparable in the way in which we have defined comparability, the selection procedure to be described will not obtain maximal results. Very probably, it would be desirable to employ weights determined by subjective judgment as to the importance of the job rather than to employ ratings or other such criteria in raw or standard score form. However, it should possibly be emphasized that the form of the procedures to be developed here are not dependent upon the character of the units employed, even though the results themselves may be considerably affected.

If we assume that our criteria are expressed in terms of dollars

as units, the object of the selection procedure would be to maximize the saving. Procedures for accomplishing this will be developed in the following presentation. Let us assume that:

1. All zero-order and partial regressions are linear.
2. All predicted criterion values have been computed for all criteria for the same battery of tests. The symbol $\bar{y}_i$ will refer to a predicted value of any given criterion i.
3. All statistical constants refer or apply to the sample of applicants.

It may be readily shown—simply by summing the various arrays—that

$$\Sigma_s \bar{y}_i = b_{yx} \Sigma_s x \,, \tag{1}$$

where the subscript s indicates that summation is within those above a point of cut *on the predictor*. If the criterion is expressed in terms of dollars saved, it can be seen that $\Sigma_s \bar{y}_i$ gives the amount saved for that selected group by the selection process.

With n assignments, the total saving in dollars (the criterion) would be

$$I = \Sigma_s \bar{y}_1 + \Sigma_s \bar{y}_2 + \cdots + \Sigma_s \bar{y}_i + \cdots + \Sigma_s \bar{y}_n \,. \tag{2}$$

I, the total saving in dollars, is the index to be maximized.

Note: The reader is reminded that, if no bias enters into the prediction, the algebraic sum of the errors of prediction approaches zero, so that $\Sigma\bar{y}$ equals Σy. Note though that $\sigma_{\bar{y}}$ equals $r_{xy}\sigma_y$ or, in terms of multiple prediction $\sigma_{\bar{y}}$ equals $R_{y.x_1x_2\ldots x_n}\sigma_y$. This latter point is significant. We have already indicated that σ_y, if determined in proper units, would increase as the "importance" of the job increases. Since in practice we must employ test scores as predictors, Σy must be calculated for those selected by the tests. This is equivalent to selecting on $\bar{y}$. Hence, the fact that $\sigma_{\bar{y}}$ is a function both of σ_y and the multiple correlation means that the accuracy with which the criterion can be predicted will have considerable effect upon the differential assignment of the applicants, since use of $\bar{y}$ instead of y values is equivalent to weighting according to the size of the correlation. That the weighting for purposes of differential prediction is properly a function of R should, in any event, be apparent. If the multiple correlation is zero for a given assignment, the test scores are completely unrelated to the criterion values and to the desirability of selecting individuals obtaining these scores. Hence, it is apparent that the assignment involved should be completely disregarded in selecting men for other assignments even though it may be far more important to obtain men high in that assignment than in the case of any remaining assignments. While the significance of this general principle is most evident in differential prediction, it has definite implications in employing a single predictor to select for several assignments.

The problem is that of defining the bounding surfaces distinguishing between the various assigned groups and differentiating the assigned groups from the unassigned in order to obtain the desired maximized value of I. While an exact general solution would involve further and highly restrictive assumptions (i.e., normality of the correlation surfaces) concerning the nature of the frequency functions and would in any event be exceedingly complex, we shall be content in the present paper with demonstrating that I will be maximized if arbitrarily determined "critical rejection scores" on the $\bar{y}$ values are employed to distinguish between assigned groups and those rejected entirely, while the differences between all possible pairs of such critical rejection scores are employed as "critical difference scores" on the "difference variables" to distinguish between the various possible assignments. It will be noted that the proposed solution requires that three separate propositions be demonstrated:

1. The desired bounding surface differentiating between the assigned group i and the rejected group is defined by a given critical rejection score on $\bar{y}_i$ (note the direct implication that the bounding surface is not curvilinear).
2. The desired bounding surface between any two assigned groups such as i and j is adequately defined by a critical difference score on the difference variable $(\bar{y}_i - \bar{y}_j)$.
3. The exact desired critical difference score on $(\bar{y}_i - \bar{y}_j)$ is the difference between the two critical rejection scores involved.

No equation for directly determining the exact critical rejection scores is to be developed although a method of successive approximations will be suggested.

With reference to all of the propositions to be proved, it will be helpful to note that it is axiomatic that I is maximized when (1) all $\bar{y}$ values for any rejected individual are equal to or lower than any $\bar{y}$ value for assigned individuals on the criterion of their assignments and (2) it is not possible to replace individuals who have higher $\bar{y}$ values on other than the criterion of their assignment with unassigned individuals such that the loss effected by replacement is smaller than the gain effected by change in assignment, or it is not possible to effect a gain by any combination of such replacements and reassignments.

Although the first of our propositions must be true if the assumption of linear regression is met, this relationship between the assumption and the proposition is not immediately evident. Suppose, in a multidimensional space with the $\bar{y}$'s as coordinates, we were to prepare separate plots of individuals having successive $\bar{y}_i$ scores from

—3.0 up to +3.0. That is, the first plot would consist of all individuals having a $\bar{y}_i$ score of —3.0, the second of individuals having a $\bar{y}_i$ score of —2.9, etc. Each of these plots would constitute a "slice" through the hyperspace in which the individuals were plotted. It is quite apparent that our rejection bounding surface, that is, our means of distinguishing between those to be rejected and those belonging in the given assignment, must be one of these "slices," since within each "slice" all $\bar{y}_i$ values are the same and within the "slices" above or below all $\bar{y}_i$ values are, respectively, larger or smaller. In other words, it makes no difference in summing the $\bar{y}_i$ values which of a group of individuals having a constant $\bar{y}_i$ score are chosen for rejection or for assignment elsewhere and, in addition, all those having such a constant score are to be preferred to those having a lower score. Hence, within any given "slice" it makes no difference in maximizing $\sum_s \bar{y}_i$ which individuals are selected for other assignments, although the relative proportions chosen from successive slices do very definitely affect $\sum_s \bar{y}_i$. Among those not assigned elsewhere, a lower value would never be selected in preference to a higher $\bar{y}$ value, no matter what scores are obtained on criteria of other assignments. To state otherwise would be to imply that $\bar{y}_i$ is not the best prediction of y_i. In other words, the fact that we are concerning ourselves with $\bar{y}_i$ instead of y_i, the criterion itself, does not influence the validity of the foregoing statement so long as the $\bar{y}_i$ values represent the "best" prediction obtainable from the test scores in terms of which the individuals are plotted. From our assumption of linear regression lines it may be stated, in any event, that the mean criterion score is equal to the mean predicted criterion score for any segment of our "slice." From this assumption it also follows that none of the "slices" is curvilinear. As these "slices" become infinitesimally thin they are defined exactly by a critical score on $\bar{y}_i$, since they consist of individuals having the same constant $\bar{y}_i$ score. Note that no statement concerning the magnitude of the critical score has been made. It is merely shown that the bounding surface is defined by some critical score on $\bar{y}_i$. Note also that the bounding surface is perpendicular to the multiple regression line determining the given $\bar{y}_i$ values. Thus we have demonstrated the first of our three propositions, namely, that the desired bounding surfaces differentiating between an assigned group i and the rejected group is defined by a given critical score. It would possibly be more appropriate to say that we have shown that it follows directly from our assumption of linearity of regression lines and have elaborated somewhat its meaning and implications.

The demonstration of the second proposition follows that of the first almost exactly. If we were to plot individuals having the same difference scores $(\bar{y}_i - \bar{y}_j)$ we would obtain a series of "slices" or

bounding surfaces analogous to those obtained by "slicing" on $\bar{y}_i$. From our assumptions of linearity of regression lines these bounding surfaces would also be linear. The necessity that each bounding surface differentiating between any two assignments must be defined by a constant value of difference variables $(\bar{y}_j - \bar{y}_k)$ is evident when it is realized that the differences between the $\bar{y}$ values are a direct indication of the gain or loss in I to be effected by shifting individuals from one assignment to another. Thus the second of the three points to be proved also follows almost directly from the assumptions of linearity of regression lines.

The next and third point to be demonstrated is that I is maximized when each critical difference score is the difference between the critical rejection scores on the $\bar{y}_i$ values of the two assignments involved, assuming that the proportion of individuals in each assignment remains constant. Suppose we consider the plane formed by plotting paired $\bar{y}_i$ and $\bar{y}_j$ values. To obtain agreement with our first proposition, assigned groups i and j must be separated from those rejected by critical scores on $\bar{y}_i$ and $\bar{y}_j$. We will refer to these critical scores as ${}_c\bar{y}_i$ and ${}_c\bar{y}_j$. The rejection boundaries defined by the critical scores would be straight lines, each perpendicular to its corresponding axis. To obtain agreement with our second proposition the two assigned groups must be separated from each other by one of a family of lines defined by constant values of the difference $(\bar{y}_i - \bar{y}_j)$. Parenthetically the reader is reminded that the slope of all members of this family of lines is the ratio between the s.d.'s of $\bar{y}_i$ and $\bar{y}_j$, while the s.d.'s are in turn equal to the multiple correlation of the several predictors for the given assignment times the s.d. of the criterion of that assignment.

Suppose that the third proposition—which remains to be demonstrated—has been assumed valid in separating the three groups and the boundary between the two assigned groups is that one of the family of difference lines identified by the difference between the two critical rejection scores or ${}_c\bar{y}_i - {}_c\bar{y}_j$. If this were the case, the three boundaries would have a common point of intersection. Since the critical difference score may be increased or decreased only, and since the critical rejection scores are determined after each such change in the critical difference score by the requirement that the number in each category remains constant, it is clear that only these two types of changes are possible. Consider Figure 1. Here, we have assumed that on the plane formed by $\bar{y}_i$ and $\bar{y}_j$ the critical rejection score for $\bar{y}_i$ is 2.0; the critical rejection score is 1.0 for $\bar{y}_j$, and the critical difference score is 1.0. These are indicated by the heavy lines. Suppose that the critical difference score were increased so as to transfer m individuals

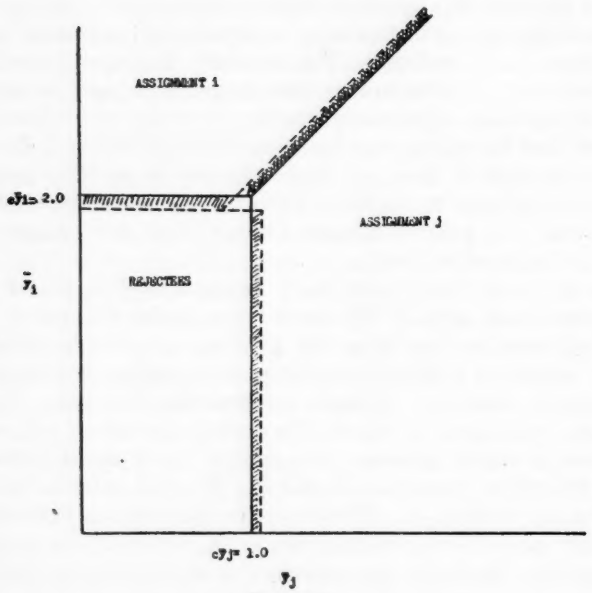

FIGURE 1

from assignment i to assignment j and changes were made in the critical rejection scores so as to allow no change between the number in assignment i and j before the transfer and the number obtained afterwards. The three shaded areas between the pairs of heavy and dotted lines delineate the individuals whose assignment is changed. It is obviously the resulting changes in the criterion of their assignment and the consequent change in the criterion scores of these individuals that will change I. In the shaded area between the two critical difference scores the m individuals involved would have been shifted from assignment i to assignment j. Since, in the case of those individuals directly on the original rejection lines (the heavy lines in Figure 1), the difference between the $\bar{y}_i$ and $\bar{y}_j$ scores was 1.0, the shift of each such individual to assignment i would decrease the over-all sum by exactly 1.0, since the criterion score for such an individual on criterion i, their assignment after the shift, is exactly 1.0 larger than their score on criterion j, or that for their previous assignment. However, the average difference in criterion scores of the individuals shifted is somewhat larger than 1.0, since the shift increased the difference critical score. Hence the decrease effected in each individual's criterion score would be 1.0 plus an increment which we

shall call Δd. The effect of this shift on the over-all sum would be to add, algebraically, $-m\ (1.0 + \Delta d)$. The individuals in the shaded area between the heavy and dotted line indicating the two critical rejection scores on $\bar{y}_j$ are being shifted to the rejected groups. Since, on the average, their $\bar{y}_j$ scores are somewhat above 1.0, this shift adds $-m(1.0 + \Delta\bar{y}_j)$ to I. By analogous reasoning, the inclusion of those in the third shaded area adds $m(\ 2.0 - \Delta\bar{y}_i)$ to I. The net change is then $-m(1.0 + \Delta d)\ -m(1.0 + \Delta\bar{y}_j)\ + m(2.0 - \Delta\bar{y}_i)$, which reduces to $m(-\Delta d - \Delta\bar{y}_i - \Delta\bar{y}_j)$, since m was made constant in order that the total in each assignment would remain the same.

Figure 2 is to be interpreted in the same manner as Figure 1. The shading indicates the areas affected by the changes in the critical difference scores and compensating changes in the critical rejection scores. In the shaded area between the two critical difference scores the mean difference is somewhat smaller than 1.0, so that in shifting individuals to assignment i, the change effected in I would be $m(1.0 - \Delta d)$. In the same way it can be seen that the shaded area between the heavy and dotted lines for $\bar{y}_i$, which represents individuals who have been shifted from assignment $\bar{y}_i$ to rejection, have an average score of something over 2.0, or say $2.0 + \Delta\bar{y}_i$. The loss occasioned by shifting to the rejected group is $-m(2.0 + \Delta\bar{y}_i)$. In-

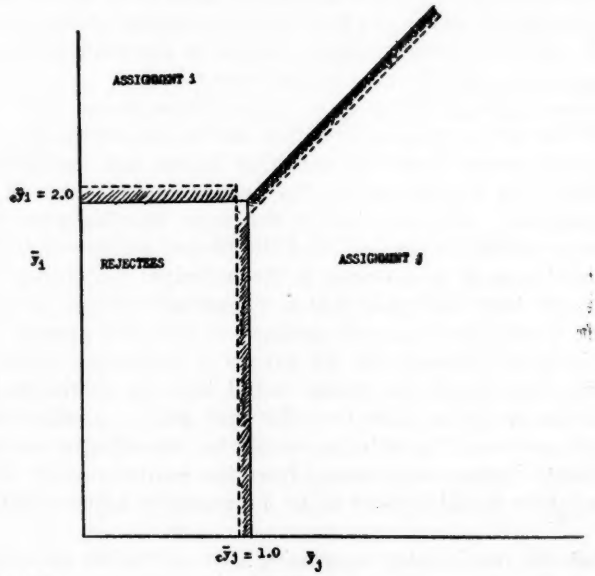

FIGURE 2

dividuals in the corresponding shaded area for $\bar{y}$; have an average criterion score of $1.0 - \Delta\bar{y}_j$ so that the effect of changing these individuals from the rejected group to assignment j is to add $m(1.0 - \Delta\bar{y}_j)$ to I. The total of these several changes is again $m(-\Delta d - \Delta y_1 - \Delta y_2)$. The area in which individuals formerly rejected are now given assignment j in Figure 1 and the area in Figure 2 where individuals formerly rejected are given assignment i have been neglected, since they are second-order differentials. It is true, of course, that the changes in the critical scores of $\bar{y}_i$ and $\bar{y}_j$ required in order to hold the proportion of cases in each assignment constant will mean that in other planes the critical difference score for, say, $\bar{y}_i$ and $\bar{y}_k$ or $\bar{y}_j$ and $\bar{y}_k$ is no longer equal to the difference between the critical scores on the two criteria defining the particular dimension in question and, it might be argued, the changes in the critical scores of other variables that will be required in order to hold the number in each assignment constant might raise I rather than lower it. However, such changes are exactly analogous to those we have discussed, and the foregoing proof will apply in the case of these further changes.

Since changes in critical scores from our proposed maximal position can be of two types only, and since it has been shown that with each of these changes a decrease in I is effected, our third proposition has been demonstrated. When all critical difference scores are equal to the differences between the two critical rejection scores, no change in critical difference scores which maintains the same proportion in each assignment will effect an improvement in I.

Demonstration of these three propositions means that with selection of any set of critical rejection scores, determination of critical difference scores from the rejection scores and assignment will be such that I is maximized for the proportions of cases in each of the assignments. The selection of the most feasible procedures of determining critical scores such that the proper number of individuals are allocated to each assignment is the principal remaining problem. It has already been indicated that a theoretical solution to this problem which would have general application does not appear possible, since assumptions concerning the nature of frequency solids formed by plotting individuals on $\bar{y}$-axes would have to correspond rather closely to the empirical data in order that such a solution would be useful. In any event the solution would be exceedingly complex and very probably rather cumbersome from the computational viewpoint. The alternative would appear to be a successive approximation procedure

A definite step-by-step successive approximation procedure will be presented. Although this procedure is not necessarily the most

rapid, it is simple and easy to apply. Let us assume that we are starting with a list of $\bar{y}$ values for all assignments and all individuals. The actual steps of the proposed procedure are as follows:

1. Prepare frequency distributions of $\bar{y}$ values for each assignment, recording the identification number of each individual instead of a frequency tally.
2. Draw a line through the top part of the frequency curve isolating, in the segmented portion, the required number of cases for that assignment. The point of cut of this line is the first estimate of the critical rejection score (designated ${}_c\bar{y}_i$).
3. Compute critical difference scores $({}_c\bar{y}_i - {}_c\bar{y}_j)$.
4. Identify in each frequency distribution, by referring the identification number of selected individuals to the listing of $\bar{y}$ values and to the critical rejection scores, all individuals above ${}_c\bar{y}$ in other frequency distributions.
5. At the time that the $\bar{y}$ values are located in the listing, refer to the critical difference scores and determine in which assignment the individual belongs. Indicate in some way on the individual's identification number in the frequency distribution that he is or is not to be included in that assignment.
6. Lower the critical scores to include enough additional subjects to replace those assigned elsewhere.
7. Determine whether any of the new assignees were included (either in the first or second selection) in any other assignment.

TABLE 1

Listing of $\bar{y}$ Values in Numerical Example

Man No.	$\bar{y}_i$ Values				Man No.	$\bar{y}_i$ Values			
	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$		$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$
1	— .8	—2.4	.1	— .4	11	— .7	—1.2	.3	.1
2	— .3	—1.9	.3	.7	12	— .9	—1.1	— .7	—1.4
3	— .4	—1.8	.1	.3	13	— .3	—1.1	.1	— .9
4	—1.1	—1.7	.6	1.1	14	— .8	—1.0	— .3	.5
5	— .1	—1.6	— .8	—1.0	15	— .4	—1.0	.6	0.0
6	— .5	—1.5	.7	— .3	16	— .3	—1.0	— .5	.3
7	— .7	—1.4	— .1	—1.3	17	.1	— .9	—1.0	2.4
8	— .6	—1.3	.2	.4	18	— .5	— .9	.8	.9
9	— .1	—1.3	— .6	— .4	19	.2	— .9	.5	— .5
10	— .2	—1.2	— .9	—1.1	20	— .2	— .8	.2	1.2

Table 1 (Continued)

Listing of $\bar{y}$ Values in Numerical Example

Man No.	$\bar{y}_i$ Values				Man No.	$\bar{y}_i$ Values			
	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$		$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$
21	— .7	— .8	.6	.9	61	— .5	.3	.4	1.0
22	.6	— .8	— .2	.9	62	— .1	.3	.3	0.0
23	.3	— .7	.2	— .1	63	— .3	.3	— .4	—1.8
24	.7	— .7	— .6	.2	64	.1	.3	.2	— .9
25	— .2	— .7	.3	.6	65	0.0	.4	— .4	— .2
26	.2	— .6	.2	.8	66	.6	.4	.1	.4
27	— .9	— .6	.1	— .3	67	1.0	.4	0.0	1.8
28	— .1	— .6	— .2	— .5	68	.4	.4	0.0	0.0
29	— .6	— .5	—1.1	—1.7	69	.6	.5	— .7	— .8
30	.5	— .5	— .5	— .9	70	.4	.5	1.1	1.3
31	.2	— .5	— .1	.2	71	— .2	.5	.1	— .8
32	— .2	— .5	0.0	1.0	72	.9	.5	0.0	— .6
33	— .3	— .4	.3	— .7	73	— .1	.6	0.0	1.2
34	— .6	— .4	.4	1.5	74	0.0	.6	.9	1.9
35	.1	— .4	.5	.3	75	.1	.6	.4	.4
36	.4	— .4	.7	.1	76	.8	.7	.6	—2.4
37	— .5	— .3	— .1	— .6	77	.1	.7	— .4	— .1
38	— .4	— .3	0.0	.4	78	0.0	.7	.1	.5
39	.7	— .3	— .1	.6	79	.1	.8	.3	1.0
40	.1	— .3	— .3	—1.0	80	0.0	.8	— .5	— .3
41	— .2	— .2	.4	1.7	81	.1	.8	— .1	0.0
42	— .3	— .2	.3	.8	82	.8	.9	0.0	— .1
43	—1.0	— .2	.2	— .2	83	— .2	.9	— .2	.2
44	— .1	— .2	.7	— .7	84	.3	.9	— .6	.5
45	— .1	— .1	— .3	— .2	85	0.0	1.0	— .6	—1.3
46	.2	— .1	— .9	— .6	86	.2	1.0	— .4	.5
47	— .1	— .1	.5	—1.0	87	.5	1.0	— .1	— .2
48	— .4	— .1	1.0	— .5	88	.4	1.1	— .2	—1.2
49	.2	0.0	— .2	— .7	89	.3	1.1	— .3	— .4
50	.3	0.0	.5	1.6	90	0.0	1.2	.8	1.3
51	.6	0.0	— .1	—1.6	91	.9	1.2	0.0	1.1
52	0.0	0.0	.2	— .5	92	0.0	1.3	— .2	—1.5
53	.3	.1	.4	— .4	93	1.1	1.3	— .1	— .1
54	.5	.1	.1	.8	94	.1	1.4	— .3	—1.9
55	.7	.1	0.0	.6	95	— .4	1.5	— .3	.3
56	.3	.1	.4	.7	96	.2	1.6	— .2	— .8
57	— .3	.2	— .3	—1.1	97	0.0	1.7	— .4	.2
58	— .6	.2	— .7	— .3	98	.3	1.8	.1	1.4
59	— .1	.2	— .8	.7	99	0.0	1.9	.1	.1
60	.4	.2	— .5	—1.2	100	.5	2.4	0.0	.1

8. If so, repeat steps 5, 6, and 7 until the required number are in each assignment and no individual is in more than one assignment.

A numerical example will serve to illustrate this procedure. Let us assume that 13, 15, 5, and 15 per cent of the population of applicants are desired for, respectively, assignments 1, 2, 3, and 4. The numerical values of the $\bar{y}$ values are given in Table 1 and the frequency distributions to be prepared as step 1 are given in Table 2. The heavy lines on the distributions indicate the first selection. The critical scores are, respectively, .6, 1.0, .8, and 1.0. This completes step 2. The critical difference scores obtained as step 3 are:

Assignment i	*Assignment j*		
	2	3	4
1	.4	.2	.0
2		—.2	—.4
3			—.2

Note that the order of the two variables in subtracting determines the sign of the difference. Assignment is made to the first of the two assignments when the value of the difference variable exceeds that of the difference score.

Identification numbers appearing in more than one frequency distribution have been primed. A line has been ruled through numbers when assignment is elsewhere (Table 2). This completes step 5.

Beyond the first approximation and reassignment, little additional labor is involved in at least this numerical example. In assignment 1, man number 67 (the only individual reassigned) was replaced by man number 30. In assignment 2, men numbers 90, 91, and 93 were reassigned and replaced by men numbers 83, 84, and 85. In assignment 3, man number 74 was reassigned and replaced by man number 90. However, after reapproximating the critical difference score between assignments 3 and 4, man number 90 was assigned to 4 (or remained there). Hence man number 6 was the eventual replacement. In assignment 4 men numbers 70, 91, and 98 were reassigned and replaced by men numbers 18, 21, and 79. However, 18 and 21 were assigned to 3 and 1, respectively, and after the critical difference scores were reapproximated they still remained there. Hence man number 26 was placed in assignment 4.

If it is desired to speed the procedure by making subjective allowances for additional factors in arriving at approximations to the critical scores, the following general principles may be helpful:

1. Assignments whose $\bar{y}$ values have relatively large standard

deviations will tend to have high critical scores, since the relative proportion of the individuals in the overlapping areas, i.e., areas above the critical scores of any pair of $\bar{y}$ values, assigned to a given category will depend upon the relative size of the s.d.'s of the $\bar{y}$ values.

2. Since degree of overlapping is a function of degree of intercorrelation, all critical scores will increase as the degree of intercorrelation of $\bar{y}$ values decreases.
3. In making approximations beyond the first, it should be remembered that when shifts are made in critical rejection scores so as to increase the number in a given category, the changes in the critical difference scores which automatically follow will decrease the number in the remaining assignment.

It is quite possible that the approximation procedures would be much more laborious if critical scores were lower and/or the size of the sample larger than in the numerical example. The author does not feel that it is feasible to offer definite information concerning the labor required in the various types of situations in which the procedures might be applied. The problem illustrated was, however, changed to the extent of requiring 25% in each assignment and the solution obtained by the author in less than one hour's time.

TABLE 2

Frequency Distributions of $\bar{y}$ Values for Four Hypothetical Assignments*

$\bar{y}$	Assignment 1										Assignment 2			
2.4											100'			
1.9											99			
1.8											98'			
1.7											97			
1.6											96			
1.5											95			
1.4											94			
1.3											92	~~93~~'		
1.2											~~90~~'	~~91~~'		
1.1	93'										88	89		
1.0	~~87~~'										85	86	87	
.9	72	91'									82	83	84	
.8	76	82									79	80	81	
.7	24	39	55								76	77	78	
.6	22	51	66	69							73	74	75	
.5	30	54	87	100							69	70	71	72
.4	36	60	68	70	88						65	66	67	68
.3	23	50	53	56	84	89	98				61	62	63	64
.2	19	26	31	46	49	96	86				57	58	59	60
.1	17	35	40	64	79	94	75	77	81		53	54	55	56
.0	74	85	90	99	78	52	65	80	92	97	49	50	51	52
- .1	5	9	28	44	45	47	59	62	73		45	46	47	48
- .2	10	20	25	32	41	83	71				41	42	43	44
- .3	2	13	16	33	42	57	63				37	38	39	40
- .4	3	15	48	38	95						33	34	35	36
- .5	6	18	37	61							29	30	31	32
- .6	8	29	34	58							26	27	28	
- .7	7	11	21								23	24	25	
- .8	14	1									20	21	22	
- .9	12	27									17	18	19	
-1.0	43										14	15	16	
-1.1	4										12	13		
-1.2											10	11		
-1.3											8	9		
-1.4											7			
-1.5											6			
-1.6											5			
-1.7											4			
-1.8											3			
-1.9											2			
-2.4											1			

* Identification numbers are listed in the place of frequency tallies.

TABLE 2 (Continued)

Frequency Distributions of $\bar{y}$ Values for Four Hypothetical Assignments*

$\bar{y}$	Assignment 3										Assignment 4			
2.4											17			
1.9											74'			
1.8											67'			
1.7											41			
1.6											50			
1.5											34			
1.4											~~98~~'			
1.3											~~96~~'	90'		
1.2											20	12		
1.1	70'										4	~~91~~'		
1.0	48										32	61	79	
.9	83	74'									18	21	22	
.8	18	90									26	42	54	
.7	6	36	44								2	56	59	
.6	4	15	21	76							25	39	55	
.5	19	35	47	50							14	78	86	84
.4	34	41	53	56	61	75					8	38	66	75
.3	2	11	25	33	42	62	79				3	16	35	95
.2	8	20	23	26	43	52	64				24	31	83	97
.1	1	3	13	27	54	98	71	66	78		11	36	100	99
.0	32	55	73	38	67	68	72	82	91	100	15	62	68	81
- .1	7	31	37	39	51	81	87	93	99		23	77	82	93
- .2	22	28	49	88	96	92	83				43	45	87	65
- .3	14	40	45	57	89	94	95				6	27	58	80
- .4	63	65	77	97	86						1	9	53	89
- .5	16	30	80	60							19	28	48	52
- .6	9	24	84	85							37	46	72	
- .7	12	58	69								33	44	49	
- .8	5	59									69	71	96	
- .9	10	46									13	30	64	
-1.0	17										5	40	47	
-1.1	29										10	57		
-1.2											60	88		
-1.3											7	85		
-1.4											12			
-1.5											92			
-1.6											51			
-1.7											29			
-1.8											63			
-1.9											94			
-2.4											76			

* Identification numbers are listed in the place of frequency tallies.

PSYCHOMETRIKA—VOL. 11, NO. 3
SEPTEMBER, 1946

THE RELATION OF MULTISERIAL ETA TO OTHER MEASURES OF CORRELATION

ROBERT J. WHERRY
UNIVERSITY OF NORTH CAROLINA
AND
ERWIN K. TAYLOR
PERSONNEL RESEARCH SECTION, A.G.O.

Ordinary product-moment correlation and regression methods are frequently not immediately applicable to qualitative data, whereas multiserial r, point-multiserial r, and multiserial eta can be easily applied. The multiserial r is rejected for prediction since it tells us only what the correlation *might be* if certain assumptions *were* true and if we *could* measure what is *not now* measured. The point-multiserial r and multiserial eta are identical when the number of categories is two but differ when it is three or greater. The multiserial eta is identical with the product-moment r when categories are assigned scale values equal to their means on the continuous variable. With three or more categories, the point-multiserial r, which *assumes* linearity with *equal* step intervals, is always lower than the multiserial eta, which *forces* linearity by adoption of *unequal* step intervals based upon difference in criterion attainment. While the multiserial eta expends one degree of freedom with the weighting of each category, this is known and correctable, whereas the vague partial loss of degrees of freedom due to the ordering of categories in the point-multiserial is not correctable.

Many problems in psychology present situations where the ordinary product-moment correlation coefficient is not immediately applicable. Suppose one wished to study the effect of dominant color or type of appeal or kind of type font on sales resulting from the use of advertisements varying only with respect to one of these variables, or to compare the importance of one such variable with another. The ordinary product-moment coefficient would not apply, but one could use a multiserial r, a point-multiserial r, or the multiserial eta coefficient to measure the degree of relationship. If actual prediction (use of regression equations) is necessary, only the point-multiserial r is usually employed. This paper will point out the superiority of the multiserial eta approach over the other methods in certain cases.

The correlation coefficient can be written in the form

$$r^2 = 1 - \frac{\sigma_{\bar{Y}}^2}{\sigma_Y^2} \qquad (1)$$

where $\sigma_{\bar{Y}}^2$ is the variance of Y about the regression line of Y on X and σ_Y^2 is the variance of Y about its own mean. The corresponding formula for eta can be written

$$\eta^2 = 1 - \frac{\sigma_{Y_c}^2}{\sigma_Y^2}, \tag{2}$$

where $\sigma_{Y_c}^2$ refers to the variance of Y as measured about the mean of each category. In order to demonstrate the relation of equation (2) to other correlation methods, it is necessary to obtain the equation in another form. We will let $X_a, X_b, \cdots, X_m$ represent the various categories of a qualitative variable. Y_i will represent the criterion (continuous variable) score of an individual in category X_i, while i will represent the *a posteriori* probability that an individual will fall in the category X_i.

From equation (2) we can then write

$$\eta^2 = 1 - \frac{\sum_a^m [\Sigma Y_i^2 - (\Sigma Y_i)^2/Ni]}{\Sigma Y^2 - (\Sigma Y)^2/N}. \tag{3}$$

Reducing both terms to a common denominator, combining similar terms, and noting that ΣY^2 equals $\Sigma_a^m(\Sigma Y_i^2)$ gives

$$\eta^2 = \frac{\sum_a^m [(\Sigma Y_i)^2/Ni] - (\Sigma Y)^2/N}{N\sigma_Y^2}. \tag{4}$$

This equation is equivalent to equation (175) in Peters and Van Voorhis (1) and is probably the best computational form when the number of categories is large. Still further modification is necessary for our purposes, however.

If we note that

$$(\Sigma Y)^2/N = [\sum_a^m (\Sigma Y_i)]^2/N, \tag{5}$$

expand (5), substitute in (4), rearrange, and factor out $1/N$, we have

$$\eta^2 = \frac{\{\sum_a^m [(\Sigma Y_i)^2/i - (\Sigma Y_i)^2] - 2\Sigma Y_a \sum_b^m Y_i - 2\Sigma Y_b \sum_c^m Y_i - \cdots - 2\Sigma Y_l \Sigma Y_m\}/N}{N\sigma_Y^2}. \tag{6}$$

Reducing the terms in each bracket to a common denominator and adding, and multiplying the numerator and denominator by N yields

$$\eta^2 = \frac{\sum_a^m \frac{1-i}{i}(\Sigma Y_i)^2 - 2\Sigma Y_a \sum_b^m Y_i - 2\Sigma Y_b \sum_c^m Y_i - \cdots - 2\Sigma Y_i \Sigma Y_m}{N^2 \sigma_Y^2}. \quad (7)$$

Substituting the equality ($iNM_i = \Sigma Y_i$) for each category sum of Y and cancelling probabilities where possible give

$$\eta^2 = \frac{N^2 \sum_a^m [i(1-i)M_i^2] - 2aM_aN^2 \sum_b^m M_i - 2bM_bN^2 \sum_c^m M_i - \cdots - 2lmN^2 M_l M_m}{N^2 \sigma_Y^2}, \quad (8)$$

where M_i is a convenient abbreviation of M_{Y_i}.

Dividing both numerator and denominator by N^2, and noting that $1 - k = (a + b + c + \cdots + m) - k$, since the sum of all of the probabilities is equal to unity, give

$$\eta^2 = \frac{\begin{array}{l} a(b + c + \cdots + m)M_a^2 + b(a + c + \cdots + m)M_b^2 + \cdots \\ + m(a + b + \cdots + l)M_m^2 - 2abM_aM_b \\ - 2acM_aM_c - \cdots - 2lmM_lM_m \end{array}}{\sigma_Y^2}. \quad (9)$$

Expanding the numerator terms and collecting terms multiplied by similar probabilities give

$$\eta^2 = \frac{\begin{array}{l} ab(M_a^2 - 2M_aM_b + M_b^2) + ac(M_a^2 - 2M_aM_c + M_c^2) + \cdots \\ + am(Ma^2 - 2M_aM_m + M_m^2) + \cdots \\ + lm(M_l^2 - 2M_lM_m + 2M_m^2) \end{array}}{\sigma_Y^2}. \quad (10)$$

Noting that each expression in parenthesis in the numerator is a perfect square, we may finally condense the equation to the form

$$\eta^2 = \frac{\begin{array}{l} ab(M_a - M_b)^2 + ac(M_a - M_c)^2 + \cdots + am(M_a - M_m)^2 \\ + \cdots + lm(M_l - M_m)^2 \end{array}}{\sigma_Y^2}. \quad (11)$$

Equation (11) provides a fairly simple and straightforward method of solution of the multiserial eta when the number of categories is small. Two special forms are presented below. If the variable has only three categories, *triserial eta* takes the form

$$\eta^2 = \frac{ab(M_a - M_b)^2 + ac(M_a - M_c)^2 + bc(M_b - M_c)^2}{\sigma_Y^2}. \quad (12)$$

If there are only two categories, the biserial eta becomes

$$\eta^2 = \frac{ab(M_a - M_b)^2}{\sigma_Y^2}, \quad (13)$$

or, if we take the square root of both sides, *biserial eta* is

$$\eta = \frac{M_a - M_b}{\sigma_Y}\sqrt{ab}, \quad (14)$$

which is an exact duplicate of the *point-biserial* correlation coefficient derived earlier by Richardson and Stalnaker (2), by assigning score values or weights of 0 and 1 to the two categories.

Richardson and Stalnaker have already pointed out the advantages of the point-biserial over the more frequently quoted and used biserial coefficient. Whereas the biserial coefficient tells us what the correlation *might be* if certain assumptions *were* true and if we *could* measure what is *not now* measured, the point biserial tells us what the effective relationship *is* if we attempt linear prediction.

Although the authors cannot cite evidence, they have seen frequent derivations of a so-called *point triserial* correlation coefficient derived by assigning score values or weights of -1, 0, and $+1$ to the three categories. These derivations result in an equation of the type

$$r = \frac{(cM_c - aM_a) - M_Y(c - a)}{\sigma_y\sqrt{(a + c) - (c - a)^2}}, \quad (15)$$

where $X_c = +1$, $X_b = 0$, and $X_a = -1$. Similarly Cyril Burt (3) has evolved the *triserial* correlation, involving the usual *assumptions* of *normality* and *linearity* and *measurability*, as

$$r = \frac{M_c - M_a}{\sigma_y} \cdot \frac{1}{\frac{Z_c}{c} + \frac{Z_a}{a}}, \quad (16)$$

where Z_i = the ordinate of the normal curve at that point of subdivision represented by the probability i. The present writers would hold that the triserial coefficient is subject to the same criticisms cited against the biserial by Richardson and Stalnaker. They do not, however, accept the *point triserial* as the most happy alternative. A method which permits *triserial eta* to be used in prediction situations will be presented as a third and preferred approach.

An adaptation of a method advanced by Wherry (4) in assigning numerical values to qualitative classifications can be shown to yield results identical with triserial eta.

If to each of several qualitative categories (a, b, c, $\cdots$, m) of a given variable, X, the mean numerical value of the criterion scores obtained by members of that category is assigned as the *weight* or *score value* of that category, we have

$$X_a = M_a, X_b = M_b, X_c = M_c, \cdots, X_m = M_m. \tag{17}$$

Given the usual Pearsonian equation for the correlation coefficient,

$$r^2 = \frac{[N\Sigma XY - \Sigma X \Sigma Y]^2}{[N\Sigma Y^2 - (\Sigma Y)^2][N\Sigma X^2 - (\Sigma X)^2]}, \tag{18}$$

and letting the X values be the mean criterion score as above, we have

$$\begin{aligned}\Sigma XY &= M_a\Sigma Y_a + M_b\Sigma Y_b + M_c\Sigma Y_c + \cdots + M_m\Sigma Y_m \\ &= aNM_a^2 + bNM_b^2 + cNM_c^2 + \cdots + mNM_m^2,\end{aligned} \tag{19}$$

$$\begin{aligned}\Sigma Y &= \Sigma Y_a + \Sigma Y_b + \Sigma Y_c + \cdots + \Sigma Y_m \\ &= aNM_a + bNM_b + cNM_c + \cdots + mNM_m,\end{aligned} \tag{20}$$

$$\Sigma X = aNM_a + bNM_b + cNM_c + \cdots + mNM_m, \tag{21}$$

$$\Sigma X^2 = aN(M_a)^2 + bN(M_b)^2 + cN(M_c)^2 + \cdots + mN(M_m)^2; \tag{22}$$

but since $\Sigma XY = \Sigma X^2$ and $\Sigma Y = \Sigma X$, the right-hand member of the denominator is equal to the squared term in the numerator and will cancel, leaving

$$r^2 = \frac{N\Sigma XY - \Sigma X\Sigma Y}{N\Sigma Y^2 - (\Sigma Y)^2}. \tag{23}$$

Substituting for the numerator term, we have

$$r^2 = \frac{\begin{array}{l}N[aNM_a^2 + bNM_b^2 + \cdots + mNM_m^2] \\ \quad - [aNM_a + bNM_b + \cdots + mNM_m]^2\end{array}}{N\Sigma Y^2 - (\Sigma Y)^2} \tag{24}$$

$$= \frac{\left\{\begin{array}{l}aN^2M_a^2 + bN^2M_b^2 + \cdots + mN^2M_m^2 \\ - (a^2N^2M_a^2 + b^2N^2M_b^2 \\ + \cdots + m^2N^2M_m^2 + 2abN^2M_aM_b + 2acN^2M_aM_c \\ + \cdots + 2lmN^2M_lM_m\end{array}\right\}}{N^2\sigma_y^2} \tag{25}$$

$$= \frac{\left\{\begin{array}{l} a(1-a)N^2M_a{}^2 + b(1-b)N^2M_b{}^2 + \cdots + m(1-m)N^2M_m{}^2 \\ -2abN^2M_aM_b - 2acN^2M_aM_c - \cdots - 2lmN^2M_lM_m \end{array}\right\}}{N^2\sigma_y{}^2}. \quad (26)$$

But since (26) is now identical with equation (8) above, it follows that computing the Pearsonian product-moment linear correlation coefficient, scoring each qualitative category in accordance with the Wherry weights, is the same as computing multiserial eta.

In the case of three categories, the computing of a point-triserial (assigning weights of -1, 0, $+1$ to the categories) assumes (a) equidistant category means for the qualitative steps and (b) linearity of relationships between the two variables. The triserial eta, if used merely as an estimate of relationship, assumes nothing, but if used along with the Wherry weights (assigning weights of M_a, M_b, and M_c to the categories), assumes (a) category means for the qualitative steps equal to the difference between the criterion attainment of the category populations, and thus (b) *forces* linearity rather than *assumes* it. It is this *forcing* of linearity which causes the Pearsonian product-moment (computed on the basis of the Wherry weights) to equal eta, since the linear regression line is the best possible regression line (assuming the qualitative classes to possess the values assigned). A further advantage of the triserial eta over the point-triserial is that in a truly qualitative variable the decision as to which category shall be labelled $+1$, which 0, and which -1 may not be at all clear until at least their *order* on some continuum is established. This would most likely be accomplished by consideration of their criterion potency (if high prediction is desired they must be so ordered). The triserial eta used simply as an indicator of relationship does not require the ordering, and if used as a linear Pearsonian correlation along with the Wherry weights for the categories, looking toward prediction, the problem of order is taken care of in the weights.

A possible criticism of the triserial eta (with the Wherry weights) is that one degree of freedom is expended with each weighting of a category. This is true but is also a partial limitation of the point-triserial as well, if the *order* is based upon relationship of the categories to the criterion (as would necessarily be the case if the categories were colors, types, or classes of appeal in such a field as advertising). Indeed the complete using of a degree of freedom per category in the triserial eta is to be preferred to the vague partial loss due to ordering in the point-triserial, since in the former the proper correction is at least known (5). Everything said concerning the triserial eta and point-triserial would apply to series involving a larger number of categories. If only two categories are involved,

however, the two methods coincide. This arises from the fact that the value of the correlation coefficient is not affected by the use of primed scores, i.e., by substituting the equality

$$X' = \frac{X - k_1}{k_2}$$

for each score. For only two categories, where $X_a = M_a$ and $X_b = M_b$, it follows that if k_1 be taken as M_a and k_2 be taken as $(M_b - M_a)$, then

$$X'_a = \frac{X_a - M_a}{M_b - M_a} = \frac{M_a - M_a}{M_b - M_a} = 0$$

and

$$X'_b = \frac{X_b - M_a}{M_b - M_a} = \frac{M_b - M_a}{M_b - M_a} = 1.$$

Thus the point-biserial is simply a special primed case of the biserial eta (with Wherry weights), if the values are assigned in the proper *order*. Indeed, in the article cited above, Wherry showed that the mean-criterion-value weights for categories could be reduced to a primed series regardless of the number of categories. The only difference is that when the number of categories is greater than two the steps between the classes cannot be reduced to equal units (such as $-1, 0, +1$) unless the differences between the weights (mean criterion values) *happen* to be equal.

REFERENCES

1. Peters, C. C. and Van Voorhis. Statistical procedures and their mathematical bases. New York: McGraw Hill, 1940, p. 315.
2. Richardson, M. W. and Stalnaker, J. M. A note on the use of biserial r in test research. *J. gen. Psychol.*, 1933, 8, 463.
3. Burt, C. Statistical problems in the evaluation of Army tests. *Psychometrika*, 1944, 9, 219-236.
4. Wherry, R. J. Maximal weighting of qualitative data. *Psychometrika*, 1944, 9, 263-266.
5. Kelley, T. L. An unbiased correlation measure. *Proc. Nat. Acad. Sci.*, 1935, 21, 554-559.

PSYCHOMETRIKA—VOL. 11, NO. 3
SEPTEMBER, 1946

DIAGRAMS FOR COMPUTING TETRACHORIC CORRELATION COEFFICIENTS FROM PERCENTAGE DIFFERENCES

SAMUEL P. HAYES, JR.

A description is given of diagrams (available separately) for computing tetrachoric correlation coefficients. The diagrams are entered with "per cent of combined groups above dividing point" and difference between groups in their per cents above the dividing point.

The tetrachoric coefficient of correlation is a very simple measure of relationship, with a precise meaning: it is the coefficient of correlation of that normal surface which exactly fits the data. Before short-cut methods of approximating its value were developed, its use was frequently prohibited by the labor expended in computing it.*

This coefficient requires the assumptions that both of the variables correlated are continuous and are distributed in accordance with the normal curve of error.† However, it may also be used in many cases where the actual distributions are somewhat skewed, since moderate skewness does not affect the coefficient greatly except where the dichotomy is extreme in one or both of the variables.

It is particularly valuable in that it may be used in cases where either or both of the variables is impossible to measure. As long as each variable can be classified into two groups, such as "passed test" and "failed test," "blond" and "brunette," or "optimistic" and "pessimistic," and if the assumption can be made that the variable underlying each classification is continuous and is distributed more or less like a normal distribution, the tetrachoric coefficient provides a measure of correlation without any true measurement of the variable itself.

If a variable is classifiable into more than two groups, and if these groups can be assumed to be parts of a continuum (if this latter assumption does not hold, tetrachoric r should not be used), these groups may be amalgamated in such a way that only two groups remain, and the tetrachoric coefficient then computed. It is often useful to make several different amalgamations, calculating the r_t for

* The full formula was developed by Karl Pearson in "On the correlation of characters not quantitatively measurable," *Philos. Trans., Series A*, 1900, vol. 195, 1-47.

† Other assumptions are homoscedasticity, rectilinearity of regressions, and normal distributions in the individual columns.

each, and then averaging the values obtained. [For example, a variable with four classes might be amalgamated as follows: (1 + 2 + 3) vs. (4), (1 + 2) vs. (3 + 4), and (1) vs. (2 + 3 + 4.)] However, the amalgamation which results in the four cell values most equal in size will yield the r_t with the smallest standard error.

The coefficient is valuable also in correlating variables that *can* be measured. In such cases, the product-moment method of correlation is naturally preferable, since it does not require the assumption of normality and since the standard error of the product-moment coefficient is usually less than half as large as the standard error of the tetrachoric coefficient. However, where the distributions are really normal, the tetrachoric coefficient has exactly the same value as the product-moment coefficient would have if the latter were calculated by using an infinite number of classes of each variable. Since most biometric or psychometric variables have distributions which are nearly enough normal to justify the use of the tetrachoric coefficient, the latter may often be used to give a quick indication of whether or not it is worth while to use the more exact product-moment method.

Its usefulness and simplicity have led to the development of a number of short-cut methods of approximation of the tetrachoric coefficient. The use of approximate values is justified on the grounds that (1) any great precision in the coefficient is ordinarily spurious, because the basic data are themselves subject to errors of measurement and sampling, and are rarely distributed in perfectly normal distributions; (2) while the full formula is extremely tedious to work out, the tetrachoric coefficient is so clear and simple in meaning that its approximation is preferable to more obscure coefficients more easily computed precisely; and (3) the standard error of tetrachoric r is fairly large,* and slight variations in the value of the coefficient are therefore of little significance.

Among the short-cut tools developed for approximating the value of the tetrachoric coefficient are tables of the volume of one quadrant of a tetrachoric table† and computing diagrams based on the volume of one quadrant of a tetrachoric table.‡

Percentage differences between groups are easily calculated and are frequently used in making direct comparisons to indicate relationship, although they do not really *measure* relationship. There is,

* This standard error can now be very readily computed by using the tables presented in Samuel P. Hayes, Jr., Tables of the standard error of the tetrachoric correlation coefficient. *Psychometrika*, 1943, 8, 193-203.

† Karl Pearson (editor). Table for statisticians and biometricians. Cambridge University Press, 1931, Part II, pp. 78.

‡ L. Cheshire, M. Saffir, and L. L. Thurstone. Computing diagrams for the tetrachoric correlation coefficient. University of Chicago Bookstore, 1933.

however, a rather close correlation between percentage differences and the tetrachoric correlation coefficient. (E.g., where one group is not more than five times as large as the other, and for an attribute manifested by 30% to 70% of the combined group, a difference of 20% between the groups indicates an r_t of approximately $\pm.30$; while a difference of 50% indicates a positive r_t of about .70 or a negative r_t between .70 and 1.00.)

Computing diagrams have therefore been developed for rapid determination of tetrachoric r from percentage differences. The only computations necessary for this determination are (1) the ratio of the smaller group to the larger; (2) the percentage of the combined groups above any pre-determined dividing point; and (3) the difference between the percentage of one group above the dividing point and the percentage of the other group above the dividing point. Three examples of these diagrams are published herewith in reduced size. A complete set of these twelve diagrams, size $8\frac{1}{2} \times 11$ inches, may be obtained without charge from the Marketing and Research Service, Dun & Bradstreet, Inc., 290 Broadway, New York 8, N. Y. Reprints of this article and of its predecessor* will also be supplied as long as the limited supply lasts.

The present computing diagrams are much more compact than those previously published. In addition, they make possible more accurate interpolation, especially in cases where either variable is divided into extremely unequal groups. Finally, they involve a minimum of calculation, and a part of this is identical with that required for short-cut calculation of the standard error of r_t.†

In order to use these diagrams, the data being correlated should be arranged so that the smaller group is on the left and the smaller row is at the top. Table 1 is arranged in this way. Note that $a + b < c + d$ and $a + c < b + d$. (Of course, if more than 50% exceeded the dividing point, the table should be set up so that the top row would contain the cases below the dividing point and the bottom row cases above the dividing point. The top row must always be smaller than

TABLE 1

Proper Arrangement of Data Preparatory to Calculating Coefficient of Tetrachoric Correlation

	Group A	*Group B*	*Total*			
Above Dividing Point	567	1000	1567	a	b	$a + b$
Below Dividing Point	189	3000	3189	c	d	$c + d$
Total	756	4000	4756	$a + c$	$b + d$	N

* Hayes, Samuel P., Jr., *op. cit.*
† Hayes, Samuel P., Jr., *op. cit.*

the bottom row and Group A smaller than Group B.) This arrangement permits the use of diagrams only one half as large as would otherwise be required. The top row of a tetrachoric table arranged as in Table 1 must next be reduced to percentages of the column totals, as follows:

TABLE 2

Determination of Percentage of Each Group Above the Dividing Point

	Group A	*Group B*		
Above Dividing Point	$\frac{567}{756} = 75\%$	$\frac{1000}{4000} = 25\%$	$\frac{a}{a+c}$	$\frac{b}{b+d}$

It is now possible to decide whether there is a *positive* or a *negative* relationship between being in the top (smaller) row and being in the left (smaller) column (in this case, between being in Group A and being above the dividing point). If the percentage in the top row of Group A is *larger* than the percentage in the top row of Group B, the correlation is *positive*, and Diagrams I through VII should be used. If the percentage in the top row of Group A is *smaller* than the percentage in the top row of Group B, the correlation is *negative*, and Diagrams I and VIII through XII should be used. (Note that a *positive* correlation between being in the top row and being in Group A is the same as a *negative* correlation between being in the top row and being in Group B. The use of cell a to indicate whether the correlation is positive or negative simply provides a reliable guide to the proper diagram.)

These diagrams are serviceable for all possible dichotomies up to those where, in either or both variables, one group is 214 times as large as the other. The choice of the particular diagram to use in any given instance depends upon the relative sizes of the two groups being compared. Table 3 gives the proper diagram to use for each ratio

TABLE 3

	Positive Correlation	*Negative Correlation*
Ratio of smaller group to larger	Diagram Number	Diagram Number
1:1	I	I
.446:1	II	VIII
.189:1	III	IX
.058:1	IV	X
.0233:1	V	XI
.00468:1	VI	XII

of one group to the other, for positive and negative coefficients of correlation.

For example, if one group is approximately the same size as the other, Diagram 1 should be used both for positive and for negative correlation; if the ratio of the smaller group to the larger is .19:1, Diagram III should be used for positive correlation and Diagram IX for negative correlation.

It is important to note that the data can always be arranged in two different ways (e.g., Table 1 may be used as illustrated or may be rearranged so that the rows become columns and the columns become rows, with the figure 567 still in the upper left cell). It frequently happens that rearrangement of the data in this way will permit the use of a diagram from which the value of r_t can more easily be obtained. It is safe practice to arrange all tables in both ways and read the value of r_t independently for each. This provides a useful check as well.

The vertical scale on each diagram gives the smaller of the two percentages into which the combined groups are divided by the dividing point used.

The horizontal scale on each diagram gives all possible percentage differences between the two groups, calculated by considering the total of each group as 100%.

The value of the tetrachoric coefficient of correlation is read from the diagram by interpolation between the curving lines, each of which at every point represents a single value for the coefficient. An illustration of this follows:

TABLE 4

Fictitious Results of Test Taken by 4756 Children and Their Parents

	Group A (Children whose parents passed test)	Group B (Children whose parents failed test)	Total
Children who passed test	567	1000	1567
Children who failed test	189	3000	3189
Total	756	4000	4756

There is clearly a positive correlation between being in Group A and passing the test. The relative sizes of the two groups are determined by dividing the smaller by the larger as follows: $\frac{756}{4000} = .189$. Diagram III should therefore be used.

Percentages need to be determined only for the top row (since that is the smaller), as follows:

$$\frac{567}{756} = 75\% \qquad \frac{1000}{4000} = 25\% \qquad \frac{1567}{4756} = 32.9\%.$$

Since the percentage difference between the two groups is 50%, and the percentage of the combined groups who passed the test is 32.9%, find in Diagram III the intersection of the vertical line at 50 and the horizontal line at 32.9 (interpolating between lines 32 and 33). This intersection falls about halfway between the line of $r_t = +.60$ and the line of $r_t = +.70$. Interpolation thus gives a correlation coefficient of about +.65.

Straight-line interpolation from one of these diagrams to the next introduces very little error, rarely more than ±.005. To illustrate this, Diagram VII was constructed. This may also be used for positive correlation wherever 16% of the combined groups are above the dividing point. The vertical scale here gives the ratios of the smaller group to the larger. The points which correspond to the first six diagrams are marked heavily, and it is clear from these that interpolation along a straight line between any two diagrams (points on this diagram) results in very little error.

The method of interpolation between diagrams is as follows:

TABLE 5

Fictitious Results of Test Taken by 2000 Children and Their Parents

	Group C (Children whose parents passed test)	Group D (Children whose parents failed test)	Total
Children who passed test	200	420	620
Children who failed test	300	1080	1380
Total	500	1500	2000

The correlation is positive and the smaller group is to the larger as 500:1500 = .333:1. This falls between Diagrams II and III. It is $\frac{.446 - .333}{.446 - .189}$ or $\frac{.113}{.257}$ or .44 of the distance from Diagram II to Diagram III.

The percentages needed are as follows: $\frac{200}{500} = 40\%$; $\frac{420}{1500} = 28\%$; $\frac{620}{2000} = 31\%$.

In Diagram III, a 12% difference at line 31% (on the vertical scale) gives an r_t of +.18. In Diagram II, it gives an r_t of +.20. Interpolating between Diagrams II and III (.44 × (.20 − .18) + .18) gives an r_t of +.19.

In many cases, it may be easier and more accurate to rearrange the data in order to reduce the necessary interpolation between diagrams. In the above example, this rearrangement would be carried out as follows:

TABLE 6

Fictitious Results of Test Taken by 2000 Children and Their Parents

	Children who passed test	Children who failed test	Total
Group C (Children whose parents passed test)	200	300	500
Group D (Children whose parents failed test)	420	1080	1500
Total	620	1380	2000

The correlation is positive and the smaller group is to the larger as 620:1380 = .449:1. This is so close to Diagram II that no interpolation between diagrams is justified.

The required percentages are: $\frac{200}{620} = 32.3\%$; $\frac{300}{1380} = 21.7\%$; $\frac{500}{2000} = 25.0\%$.

Diagram II, at the intersection of lines corresponding to a percentage difference of 10.6% and to a percentage of 25% above the dividing point, gives an r_t of +.19.

It should be stressed that, when obtaining tetrachoric r by methods of approximation, computations should not be carried out beyond tenths of one per cent. The value of the coefficient should never be considered beyond the second place, and where there are extreme dichotomies (one group less than .05 of the other) or abnormal distributions, only the first place should be considered significant.

Finally, before the tetrachoric coefficient is used, there should be a very clear understanding of the assumptions it involves and the types of data with which its use is valid. Extended discussion of the assumptions its use implies and of the particular kinds of problems for the analysis of which it is a valuable tool, will be found in Pearson's original article (cited in footnote on page 163), as well as in certain statistical texts.*

* Cf. Truman L. Kelley, Statistical method. New York: Macmillan, 1924, pp. 253-258; and Charles C. Peters and Walter R. Van Voorhis, Statistical procedures and their mathematical bases. New York: McGraw-Hill, 1940, pp. 366-384.

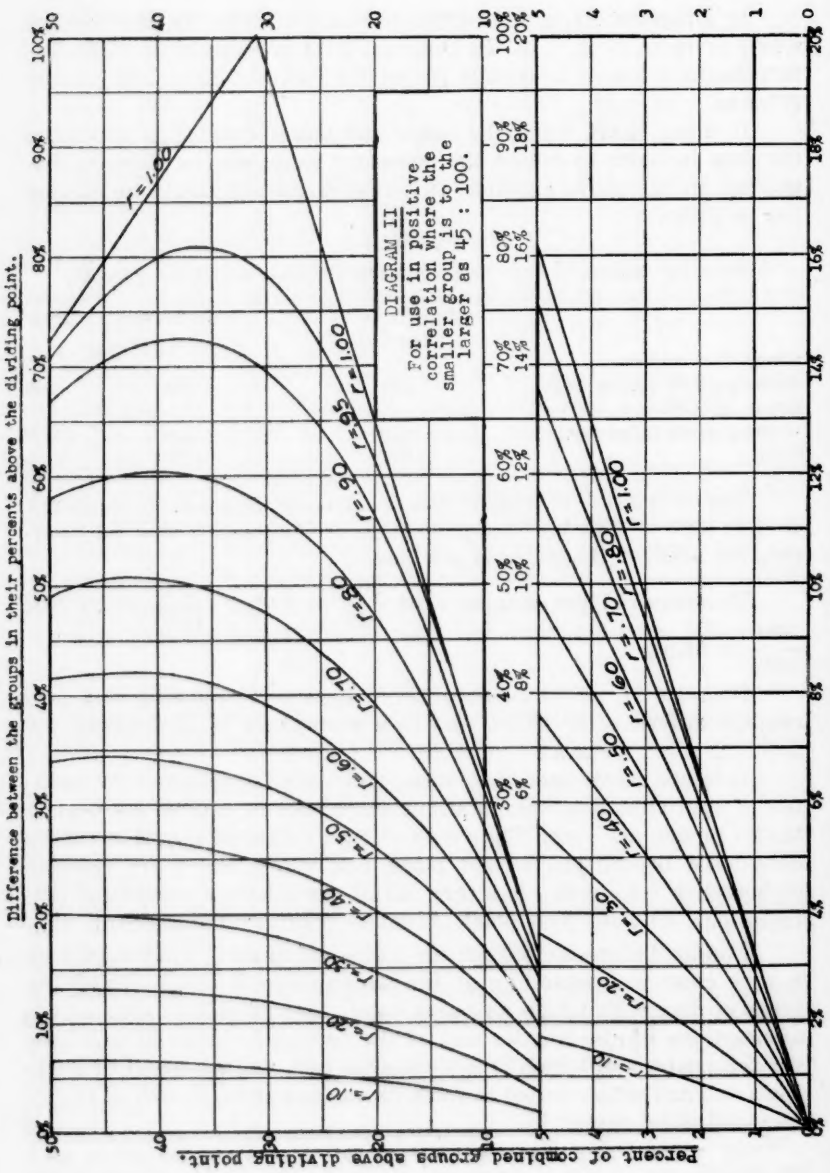
DIAGRAM II
For use in positive correlation where the smaller group is to the larger as 45 : 100.
Difference between the groups in their percents above the dividing point.
Percent of combined groups above dividing point.
r=1.00
r=.95
r=.90
r=.80
r=.70
r=.60
r=.50
r=.40
r=.30
r=.20
r=.10

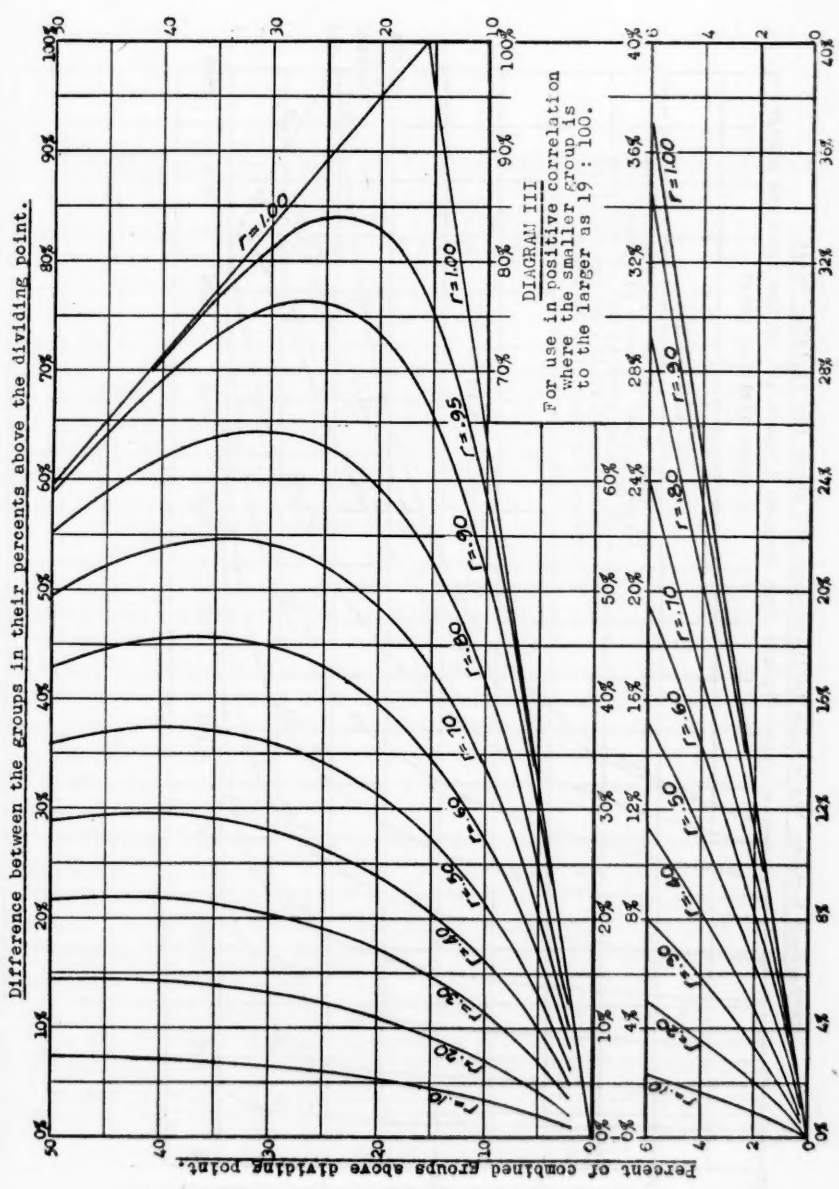
DIAGRAM III
For use in positive correlation where the smaller group is to the larger as 1.9 : 100.
Difference between the groups in their percents above the dividing point.
Percent of combined groups above dividing point.
r=1.00
r=.95
r=.90
r=.80
r=.70
r=.60
r=.50
r=.40
r=.30
r=.20
r=.10
r=0

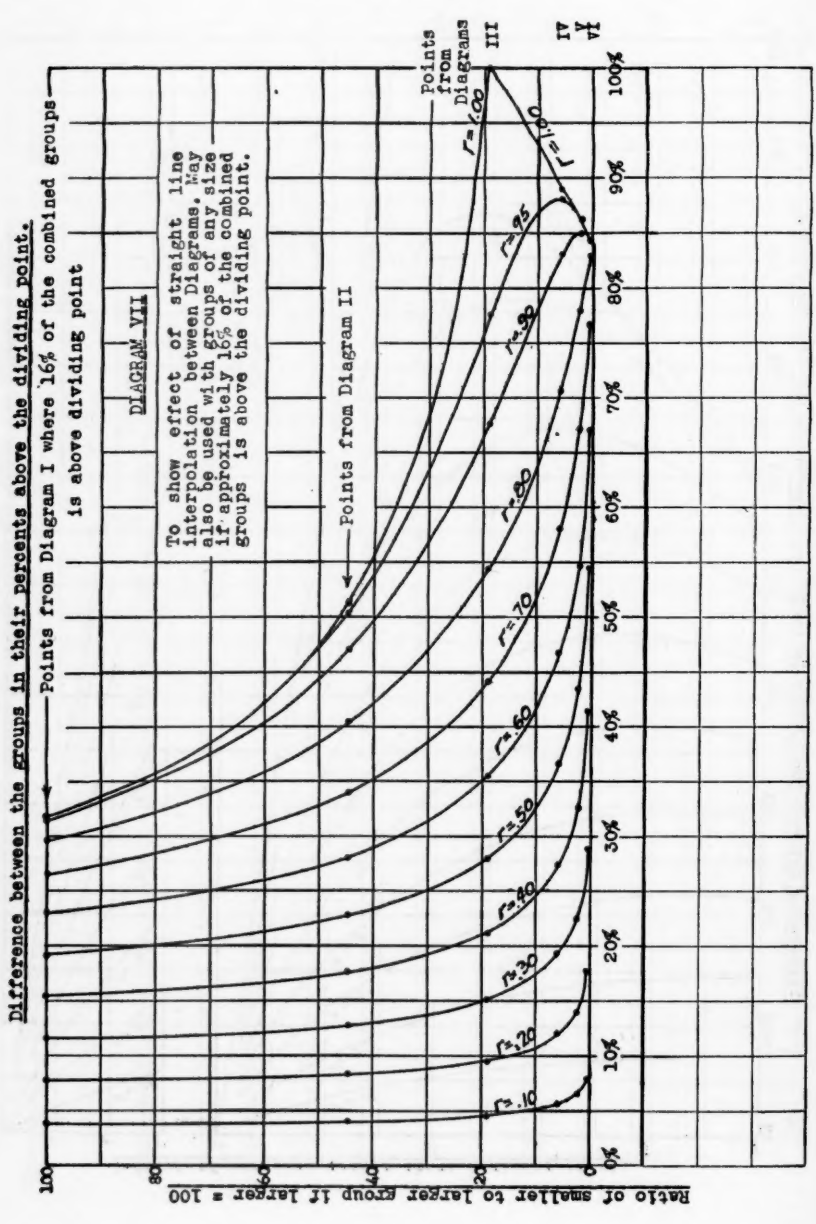
DIAGRAM VII
To show effect of straight line interpolation between Diagrams. May also be used with groups of any size if approximately 16% of the combined groups is above the dividing point.
Points from Diagram I where 16% of the combined groups is above dividing point
Points from Diagram II
Points from Diagrams III IV VI
r=1.00
r=.95
r=.90
r=.80
r=.70
r=.60
r=.50
r=.40
r=.30
r=.20
r=.10
Difference between the groups in their percents above the dividing point.
Ratio of smaller to larger group if larger = 100
100
80
60
40
20
0
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%

PSYCHOMETRIKA—VOL. 11, NO. 3
SEPTEMBER, 1946

TEST SELECTION WITH INTEGRAL GROSS SCORE WEIGHTS

ROBERT J. WHERRY
UNIVERSITY OF NORTH CAROLINA
AND
RICHARD H. GAYLORD
OHIO STATE UNIVERSITY

A new method of test selection, which attempts to combine the merits of the Toops *L*-Method with those of the Wherry-Doolittle Method, is presented. It results in integral (unit if desired) positive and/or negative (optional) weights. This flexibility makes the method applicable to all kinds of material and for both selecting items for tests and tests for batteries. An explicit solution of one test construction problem is presented. Necessary changes in method for the solution of five other types of test construction problem are presented. A few cautions are provided for potential users.

In 1941, Toops (1) presented a method designed to "obtain the maximum validity of the minimally sized test, to be scored by adding gross scores of the several sub-parts." An elaborate "theoretical" solution involving all possible combinations of all parts was claimed to be practical only when the number of those to-be-considered parts was quite small. As a "practical" solution Toops proposed a cumulative iteration procedure based on the adding of one variable at a time, with positive unit gross score weights, to a previously chosen "best" battery.

Earlier, Wherry (2) proposed a similar cumulative iteration procedure which, however, used the true regression weights and resulted, therefore, in a true multiple rather than in a simple correlation of sums. The Wherry method like that of Toops added variables one at a time to a previously chosen "best" battery rather than attempting all possible combinations.

The Toops method, designed primarily for selecting items for a test, and the Wherry method, designed primarily for adding tests to a battery, each have certain advantages and certain disadvantages. The present series of modifications is an attempt to combine the merits of the two approaches. A brief discussion of each of these methods in comparison with the present approach follows:

The Toops L-Method: Uses easy to apply positive unit weights which, however, restrict the size of the correlation obtained and pro-

hibit the application of the method to personality and interest test construction where negative weights are apt to be needed. Not particularly useful for combining tests into a battery since it permits no differential weighting. Requires a new set of tables for each variable and entries are selected by means of an elaborate celluloid strip method. No back solution is required. No exact criterion of when to stop working is provided. A method for bringing the "practical" solution somewhat in line with the "theoretical" was suggested, but would prove quite cumbersome in practice.

The Wherry-Doolittle Method: Results in hard to apply non-integral weights with no provision for keeping them all positive if desired, but does result in true multiple. Is applicable to all types of material, but is not suitable for item selection since it does not provide unit or even small integral weights. Uses a single set of cumulative tables, but does require a back solution to obtain the weights. An exact criterion for stopping the addition of further items is provided. There is no provision for bringing the "practical" solution into line with the "theoretical."

The Present Method: Results in integral (unit if desired) positive or negative (optional) weights. Has great flexibility, making it adaptable to all kinds of materials and both for selecting items for tests and tests for batteries. Uses a single set of cumulative tables, and requires no back solution for the weights. No exact criterion as to when to stop is provided. Some of the variations to be presented are already self-corrective in part, and a plan, still cumbersome but involving no new methods or formulas, is advanced for bringing the "practical" solution into line with the "theoretical."

The concept of "*L*," for which the Toops method was named and which is to be used in the present method, may be unfamiliar to some of the readers. The "*L*" occurs in many statistical formulas. While it can quickly be transformed into desired parameters, it has no decimals and involves no rounding errors due to dropping or manipulating of decimals. The "*L*" always has two subscripts (L_{ii} or L_{jj}) which refer to the variables involved. Thus if the L is relating a variable to itself (measuring variance) the subscripts are alike and the formula is

$$L_{ii} = N\Sigma X_i^2 - (\Sigma X_i)(\Sigma X_i), \tag{1}$$

being equal to N^2 times the variance. If L is relating two different variables (is measuring covariance) the subscripts are different and the equation is

$$L_{ij} = N\Sigma X_i X_j - (\Sigma X_i)(\Sigma X_j), \tag{2}$$

being equal to N^2 times the product of the square roots of the two variances times the correlation coefficient. The variables in both cases may be either simple or complex, i.e., i or j may be composite variables.

The basic equation upon which the Toops *L*-Method rests is

$$V^2/Z = L_{00}r_0^{2}{}_{(C_n+U)} = \frac{[L_{0C_n} + L_{0U}]^2}{L_{C_nC_n} + (L_{UU} + 2L_{C_nU})}, \tag{3}$$

where

C_n = composite of n items $= a + b + c + \cdots + n$,
0 = the criterion variable,
$a, b, c, \cdots, n$ = predictor variables already in the composite,
n = the last predictor variable added to the composite,
U = any predictor variable still unused,
$L_{0C_n} = L_{0a} + L_{0b} + L_{0c} + \cdots + L_{0n}$,
$L_{C_nC_n} = L_{aa} + L_{bb} + \cdots + L_{nn} + 2(L_{ab} + L_{ac} + \cdots + L_{mn})$,

and

$$L_{UC_n} = L_{Ua} + L_{Ub} + L_{Uc} + \cdots + L_{Un}.$$

The formula above stated and used by Toops is not in the best form to expedite the needed iterative calculations. Using the composite of the previous selected items, with the criterion in the numerator and with itself in the denominator, as a point of departure, the equation as stated demands the addition of a number of variable quantities to the numerator and denominator at each step, thus requiring a new set of tables for each variable added and a considerable search (thus the celluloid strips) for the variable quantities to be added.

Actually, to facilitate the needed iterative operations, the equation should use the combination of U (each unused item) and the previous composite, with the criterion in the numerator and with itself in the denominator, as the point of departure, and to these previously determined combinations should be added the increments of the n-th item just added to the battery. This would make the formula read

$$V^2/Z = L_{00}r_0^{2}{}_{(U+C_m+n)} = \frac{[L_{0(U+C_m)} + L_{0n}]^2}{L_{(U+C_m)(U+C_m)} + 2L_{n(U+C_m)} + L_{nn}}. \tag{4}$$

It is upon this rewritten formula that the present modifications of the Toops *L*-Method are based.

Two further slight modifications are necessary to put this equation into its most expeditious form. The middle term of the denominator can be split into two parts, a constant and a variable; thus:

$$2L_{n(U+C_m)} = 2L_{nC_m} + 2L_{Un}, \tag{5}$$

which would be sufficient if the original problem proposed by Toops (our Case 1) were the only consideration. To generalize the formula

so that many other problems (our Cases 2 through 6) can also be solved it is necessary only to replace the symbol U, any unused test or item, by i, the symbol for any item (used or not) and either direct (+) or reflected (−). If at the same time we expand the equation to show exactly what is added at each step in the process the equation becomes

$$V^2/Z = L_{00} r_0^{2}{}_{(i+a+b+c+\ldots+n)} =$$

$$\frac{[L_{0i} + L_{0a} + L_{0b} + \cdots + L_{0n}]^2}{L_{ii} + [(L_{aa}) + 2L_{ai}] + [(L_{bb} + 2L_{ab}) + 2L_{bi}] + \cdots + [(L_{nn} + \sum_{a}^{m} 2L_{nx}) + 2L_{ni}]} \tag{6}$$

As indicated earlier, the method requires the construction of three cumulative tables: one for V, the numerator term before squaring; one for Z, the denominator term; and one for V^2/Z, the quantity to be maximized by the addition of each new variable.

The values in the V table are started by copying all L_{0i} values in a column. To these values, and to the resulting cumulants at each step, the constant value L_{0x}, where x is the just chosen variable, is added. The new sum is then squared to obtain V^2 for the V^2/Z table.

The Z table is started by copying each of the L_{ii} values, taken from the diagonal of the L table (or computed separately), into the first column. To these values, and to the resulting cumulants at each step, is added the generalized increment of

$$(L_{nn} + \sum_{a}^{m} 2\,L_{nx}) + 2\,L_{ni}, \tag{7}$$

where x is the just chosen variable. The first part (in parentheses) of this increment is a constant added to each item, while the latter part is a variable quantity which can be copied directly from the n-th column of the $2L_{ij}$ table (to be discussed later) if the items are arranged in the same order in both tables. The first or constant part of the increment is also easily found: it consists of the original entry, L_{nn}, plus the sum of all of the quantities added as *variable* increments due to the addition of previous variables, and it can be obtained directly from the row assigned to the n-th row of the Z table.

All of the foregoing should become much clearer after the reader has seen the method of operation. Six Cases will be presented and an actual example of Case 2 will be worked out, with brief statements as to necessary changes to accomplish the remaining cases. The six problems to be solved are:

Case 1. All Positive, Unit Weights (the Toops problem),

Case 2. Positive or Negative Unit Weights,

Case 3. All Positive Integral Weights,

Case 4. Positive or Negative Integral Weights,

Case 5. Self-Corrective All Positive Unit Weights,

and

Case 6. Self-Corrective All Positive Integral Weights.

Suppose three predictors (X_1, X_2, and X_3) and a criterion (X_0) to have the sums, sums of squares, and sums of cross products, as given in Table 1, for their gross score values. Further suppose the L's to have been calculated by means of equations (1) and (2) at the beginning of the paper, which would yield the values given in Table 2. In actual practice, other than in setting up the first columns of the V and Z tables, the quantities needed will be $2L_{ij}$ and if negative weights are to be permitted, as in Case 2, we will also need the values of $2L_{i(-j)}$ and $2L_{(-i)(-j)}$, the two latter values being obtained from the former by changing signs once or twice, respectively (this latter double change leaves the sign unchanged). These double-L values for the positive and negative representation of the variable are found in Table 3.

TABLE 1

Sums, Sums of Squares, and Sums of Cross Products for Raw Scores (Assumed) $N = 10$.

	Sums of Squares and Cross Products			
	X_1	X_2	X_3	X_0
X_1	385	315	300	715
X_2	315	385	302	783
X_3	300	302	385	519
X_0	715	783	519	1762
		Sums		
	55	55	55	110

TABLE 2

The L Table: Values of L_{ii} and L_{ij} for Data in Table 1.

L	1	2	3	0
1	825	125	—25	1100
2	125	825	—5	1780
3	—25	—5	825	—860
0	1100	1780	—860	5520

TABLE 3

Values of $2L_{ij}$ for Each Variable Considered as Either Positive or Negative

$2L$	$+X_1$	$+X_2$	$+X_3$	$-X_1$	$-X_2$	$-X_3$
$+X_1$	1650	250	—50	—1650	—250	50
$+X_2$	250	1650	—10	—250	—1650	10
$+X_3$	—50	—10	1650	50	10	—1650
$-X_1$	—1650	—250	50	1650	250	—50
$-X_2$	—250	—1650	10	250	1650	—10
$-X_3$	50	10	1650	—50	—10	1650

The use of an example based upon 10 cases does not imply that actual studies should be done on such minute samples. The data here were obtained by assigning errors to an exact equation as will come out later in the analysis.

Tables similar to the above will be constructed for any problem undertaken. It should be noted, however, that it is not necessary to build these tables in advance completely. On the contrary, all that is necessary to start computation are the L_{0i} and the L_{ii} (diagonal) entries of Table 2. The various L_{ij} values of Table 2 and the $2L_{ij}$ values of Table 3 need not be calculated for any given column j until after that variable has been selected for inclusion in the test or battery. One of the main advantages of the method is that, in case of item analysis or test selection when the number of variables is large, it is not necessary to compute all inter-L's in advance.

Case 2, with Positive or Negative Unit Weights, permits each item to enter the battery only once and with either a weight of +1 or —1. Tables 4, 5, and 6 contain all of the necessary calculations (the V, Z, and V^2/Z tables) and the numbered steps give a detailed job analysis of the process.

TABLE 4

V Table for Case 2, Using Data from Tables 1, 2, and 3.

	(V_{c_a})	(V_{c_b})	(V_{c_c})	
	$L_{0i} + L_{0a} = L_{0(i+c_a)}$	$+ L_{0b} = L_{0(i+c_b)}$		
$b = +X_1$	1100 + 1780 = 2880			$a = +X_2$
$a = +X_2$	1780			$b = +X_1$
$+X_3$	—860 + 1780 = 920	+ 1100 = 2020		$c = -X_3$
$-X_1$	—1100 + 1780 = 680			
$-X_2$	—1780			
$c = -X_3$	860 + 1780 = 2640	+ 1100 = 3740		

TABLE 5

Z Table for Case 2, Using Data from Tables 1, 2, and 3.

	$L_{ii} + L_{aa} + 2L_{ia} = L_{(i+c_a)(i+c_a)}$		$+ (L_{bb} + 2L_{ab}) + 2L_{ib} = L_{(i+c_b)(i+c_b)}$
$b = +X_1$	825 + 825 + 250 = 1900		
$a = +X_2$	825		
$+X_3$	825 + 825 − 10 = 1640	+ 1075 − 50 =	2665
$-X_1$	825 + 825 − 250 = 1400		
$-X_2$			
$c = -X_3$	825 + 825 + 10 = 1600	+ 1075 + 50 =	2785

TABLE 6

V^2/Z Table for Case 2, Using Data from Tables 4 and 5.

	$+X_1$	$+X_2$	$+X_3$	$-X_1$	$-X_2$	$-X_3$
$V_{C_a}{}^2/Z_{C_a}$, if $a = X_i$	1466.7	*3840.5*	(—)896.5	(—)1466.7	(—)3840.5	896.5
$V_{C_b}{}^2/Z_{C_b}$, if $b = X_i$	*4365.5*		516.1	330.3		4198.6
$V_{C_c}{}^2/Z_{C_c}$, if $c = X_i$			1531.1			*5022.2*

And from equation (6) it follows that

$$R^2_{X_0 \cdot (X_1+X_2-X_3)} = 5022.2/5520 = .909873.$$

Job-Analysis

Step 1. Set up V, Z, and V^2/Z tables with headings as indicated, copy L_{0i} values from Table 2 for positive variables and *reverse signs* of entries for negative variables. Copy L_{ii} entries from diagonals of Table 2 for positive variables and repeat with *same signs* for negative variables.

Step 2. Compute $V_{C_a}{}^2/Z_{C_a}$ for each variable and enter in the first row of the V^2/Z table. Underline the highest positive value (all V^2/Z values are actually positive of course but consider only those items with positive V values). Draw a wavy line through the remainder of the *two* columns (the plus *and* minus columns) for the variable selected.

Step 3. Note which variable has been selected and note it as variable a to the right of Table 4, and at the left of both Tables 4 and 5. Draw wavy lines through the remainder of the row for this variable (*and* for its negative or positive counterpart) in *both* tables.

Step 4. Note the value of L_{0i} for the variable just selected and enter this value as L_{0a} in all remaining rows of the V Table (Table 4). Next add the L_{0i} and L_{0a} entries to obtain $L_{0(i+c_a)}$ entries for the next column of this table. These values will be the ones used to compute V^2 for the next row in Table 6.

Step 5. Note the entry opposite selected variable in the L_{ii} column of the Z table and copy that value in all remaining rows of the L_{aa} column of the Z table. To secure the values of $2L_{ia}$ for the next column turn to the a

column of Table 3 and copy those values in the appropriate rows of the $2L_{ia}$ column of the Z table. Add the three entries (L_{ii}, L_{aa}, and $2L_{ia}$) in each line of the Z table to obtain the values of $L_{(i+c_a)(i+c_a)}$ and record these sums in the next column of the Z tables. These will be the new Z values used to compute V^2/Z for the next row of Table 6.

Step 6. Compute the values of $V_{c_b}{}^2/Z_{c_b}$ and enter on the next row of Table 6. Underline the highest value (based on a positive V value), accepting this item as item b if it is larger than the underlined entry in the row above. (If it is lower than the underlined entry in the row above the battery is maximized and work stops). If it is accepted draw wavy lines under columns for the item and for its negative (or positive) counterpart.

Step 7. List the just selected variable as b at the right of Table 4 and at the left of both Tables 4 and 5. Draw wavy lines through the remainder of the rows for variables b and $-b$ in the two tables.

Step 8. Note the L_{0i} value of the variable just selected and enter this value as L_{0b} in the remaining rows of that column in Table 4. Next add the two last entries (columns $L_{0(i+c_a)}$ and L_{0b}) to obtain $L_{0(i+c_b)}$. These values will be used to compute V^2 for the next row of Table 6.

Step 9. Turn to the Z Table (Table 5) and in the b row, find and add the entry in the L_{ii} column and in *all* columns headed $2L_{ix}$ (only $2L_{ia}$ at this stage) and place this *constant sum* in all vacant rows of the next column of this table. Next turn to the b column of Table 3 and copy the values of $2L_{ib}$ in the appropriate cells of the next column of the Z table. For each remaining row of the Z table now add the last three entries (to the right of the last equal sign) and enter these sums in the $L_{(i+c_b)(i+c_b)}$ column of the table. These are the new Z values for computing the V^2/Z entries for the next row in Table 6.

Steps Beyond This Point. Continue to repeat steps 6-9 until the new underlined entry in step 6 is smaller than this previously underlined entry, *or* until all variables have been exhausted (as happens in next step of problem in text), *or* until increments are so small as to be insignificant or are likely results of fitting errors of measurement rather than actual increases in validity.

It will be noted that the resulting correlation from use of the combining formula based upon Case 2 has reached the very satisfactory value of .954. This values contrasts favorably with the correlation of .889 which can be attained by the best all positive unit weight combination using Case 1. The result is of course not so good as the Case 4 solution which would yield an r of 1.000, since the actual equation upon which the data were based was $X_0 = X_1 + 2X_2 - X_3$. It is suggested that the reader obtain the indicated correlation coefficients for Cases 1 and 4 to make sure that he understands the

method. In order to solve the remaining cases certain changes in procedure are necessary. These changes are:

Case 1. Each item is allowed to have a positive weight of unity or a weight of zero. The directions for Case 2 above can be *simplified* by dropping out all references to negative variables in Tables 3, 4, 5, and 6 and in the directions themselves. Otherwise the method is unchanged.

Case 3. Here each item can have a weight of either zero or any positive integer such as 1, 2, 3, 4, etc. As in Case 1 all reference to or inclusion of negative items is omitted. The only other change in directions is to *not* draw the wavy lines referred to in steps 2, 3, 6, and 7. This means that although a variable has already been admitted to the battery it still remains as a possibility for further admissions. The number of times it is admitted constitutes its *weight.* This case is suggested as useful for public or industrial battery building where negative weights might be considered undesirable from the public relations viewpoint.

Case 4. Here each item can have a weight of zero, of any positive integer (1, 2, 3, 4, etc.), or of any negative integer (—1, —2, —3, etc.). So long as the integers do not become excessive in size, this case will quickly yield results approximating the true multiple correlation. Since weights are usually rounded to such integers in any case, the method has the advantage of showing directly the validity of the battery as it will be applied. In Case 4, as in Case 2, the negative variables are used, but, as in Case 3, the wavy lines in steps 2, 3, 6, and 7 are *not* used.

Case 5. This is like Case 1 in end result but seeks to provide a means of bringing the "practical" solution in line with the "theoretical" solution. The tables would originally be set up as in Case 2 but would involve the following changes in directions:

a. the non-selection of negatively represented (reflected) items, even though they yielded the highest value in Table 6, until *after* they had been "*alerted,*" by the adoption of their positive counterpart. The column of an *alerted* item would be marked with an *X* or kept in red pencil thereafter in Table 6, until or unless it were subsequently chosen.

b. the drawing of wavy lines only after selected items, positive or alerted negative, and *not* after the counterpart of selected positive items.

If an alerted negative item were adopted for selection, this would indicate that due to other variables added to the battery the beta weight of its positive counterpart had shifted from positive to negative, which does happen *infrequently.* Even though no negative variable were

ever selected directly, the test constructor could, after top prediction was apparently reached, try systematically the arbitray selection of each alerted negative item to see if the loss engendered by its adoption could be more than overcome by the addition of one or more as yet unused positive items. In case no such restoration took place, all sections of work beginning with the arbitrary adoption of the alerted negative could be crossed out, and another alerted negative chosen in turn. The authors are frankly not very hopeful of the practical untility of this case, but it does meet an objection raised against test selection based upon adding to previously best batteries.

Case 6. This case does for Case 3 what Case 5 did for Case 1. It too would be set up like Case 2 using all variables, but would involve the following changes in directions:

a. the non-selection of negative items until after they had been "alerted" as in Case 5. Here, since the positive counterpart can be selected several times, it is necessary to have several degrees of "alertedness," which could be shown by multiple X's at the head of negative item columns in Table 6 or by the use of a succession of colors.

b. All use of wavy lines would be dropped. Positive items would never suffer restriction, but negative items adopted for inclusion would lose one degree of "alertedness" (one X would be crossed out or the item would be reduced to a lower color) each time they were selected.

This case may actually prove valuable since beta and thus gross score weights are known to shift in value as additional variables are added and this shift can be lower as well as higher. It is not expected that sizable increases in final validity will result, however.

Several words of caution to potential users of these cases and this method occur to the authors and are offered for what they may be worth:

a. In the use of cases 1 and 2, it is likely that frequently very short batteries of items will yield "maximum" validity. When items have finally paralleled the factor pattern in the criterion any effort to upset this pattern may result in lowered validity. Still the battery or test may be so short as to have relatively low reliability. It is suggested that after one "maximal" battery has been selected the selected items be dropped and the problem restarted so that a second, a third, a fourth, etc., battery is selected. This could be kept up until the final validity of the selected batteries began to fall sharply. The test builder could then compute the multiples due to successive cumulation of batteries until the gain due to added reliability was offset by the lowered validity. This last step should probably be carried out on an independent sample so that the final outcome would already have been partly cross-validated.

b. In the use of any of the methods, it may not be wise to stop work at the first sign of falling off of validity, since often a temporary offset may be more than regained by adding still more items (if the

offender acts as a suppressor for as yet unadded variables).

c. When Case 3 or 4 is used, and the suggestion in (b) above is followed, care must be taken not to get into a circular trap. For example, if $X_1 - 2X_2 - X_3$ is maximal it is true that $2X_1 - 4X_2 - 2X_3$ will give an equally good but no better result. Watch for and do not be falsely encouraged by repetition of pattern.

REFERENCES

1. Toops, Herbert A. The *L*-Method. *Psychometrika*, 1941, 6, 249-266.
2. Stead, Shartle, and Associates. Occupational counseling techniques. (Appendix V, pp. 245-252), American Book Co., 1940.

PSYCHOMETRIKA—VOL. 11, NO. 3
SEPTEMBER, 1946

NOTE ON A REANALYSIS OF DAVIS' READING TESTS

L. L. THURSTONE
THE UNIVERSITY OF CHICAGO

Correlation data on nine reading tests originally analyzed by Frederick B. Davis by the principal axes method are reanalyzed by Spearman's uni-dimensional method. It is concluded that a single common factor (reading ability) accounts for the correlations among the tests with residuals remarkably small in view of the fact that the tests were designed to test nine supposedly different skills. Three of the tests showed additional specific variance not attributable to the common factor.

Factor analysis has been the subject of much controversy ever since Spearman's famous paper of 1904. About fifteen years ago the uni-dimensional methods of Spearman were extended to the *n*-dimensional case, and discussion then turned into new forms of controversy which are still current. The purpose of this note is merely to call attention to an alternative interpretation of a recent multi-dimensional factorial study. The alternative solution is in this case relatively simple in that the uni-dimensional method of Spearman seems to be applicable instead of the more elaborate multi-dimensional solution. It is the exception to find a factor problem that can be solved so simply.

In a recent paper Davis* described a factorial analysis of nine reading tests that were designed to appraise as many different reading skills. These had been classified by experts in reading. Davis presented the correlations of the nine reading tests, and they are reproduced here as Table 1. The list of nine reading skills that were appraised by the nine tests were listed by Davis as follows:

1. Knowledge of word meanings
2. Ability to select the appropriate meaning for a word or phrase in the light of its particular contextual setting
3. Ability to follow the organization of a passage and to identify antecedents and references in it
4. Ability to select the main thought of a passage

* Davis, Frederick B. Fundamental factors of comprehension in reading. *Psychometrika*, 1944, 9, 185-197.

5. Ability to answer questions that are specifically answered in a passage
6. Ability to answer questions that are answered in a passage but not in words in which the question is asked
7. Ability to draw inferences from a passage about its contents
8. Ability to recognize the literary devices used in a passage and to determine its tone and mood
9. Ability to determine a writer's purpose, intent, and point of view, i.e., to draw inferences about a writer

Davis made a principal axes solution with nine factors; hence his interpretation has as many factors as there are tests. To use the principal axes solution imposes the condition that the nine factors shall be orthogonal, i.e., uncorrelated. From the point of view of psychology and education, it is a curious restriction to impose the condition that all of the nine abilities in this list of reading functions must be uncorrelated. Furthermore, the principal axes solution automatically imposes this restriction, no matter who the experimental subjects may happen to be.

Davis made his principal axes solution with covariances rather than with unit variances. This is a defensible procedure especially when the metric for all the variables is the same, as would be the case, for example, if all of the scores were measures of time. But here Davis was dealing with scores in nine reading tests. He made the principal axes analysis on the "variances and covariances in which the initial test variances are weighted to correspond with their relative importance in the process of reading, as determined by the pooled judgment of authorities." The first test must have been considered to be most important, with a variance of 134.70, while the fourth test was evidently considered not so important, with a variance of only

TABLE 1
Correlations of Nine Reading Tests

	1	2	3	4	5	6	7	8	9
1	...	.72	.41	.28	.52	.71	.68	.51	.68
2	.72	...	.34	.36	.53	.71	.68	.52	.68
3	.41	.34	...	.16	.34	.43	.42	.28	.41
4	.28	.36	.16	...	.30	.36	.35	.29	.36
5	.52	.53	.34	.30	...	.64	.55	.45	.55
6	.71	.71	.43	.36	.64	...	.76	.57	.76
7	.68	.68	.42	.35	.55	.76	...	.59	.68
8	.51	.52	.28	.29	.45	.57	.59	...	.58
9	.68	.68	.41	.36	.55	.76	.68	.58	...

1.22. As could be expected, the factor loadings on the first principal component vary directly with these assigned variances of the nine tests are shown in his Table 5. It seems to the writer that the problem of discovering the underlying factors in reading is obscured by first imposing weights on the test scores according to "the judgment of authorities" who are supposedly trying to find out just what the underlying factors are and how important they are.

TABLE 2
Single-Factor Solution

Tests	a_{j1}	h_j^2	r_{jj}	$(r_{jj} - h_j^2)$
1	.803	.645	.90	.25
2	.810	.656	.56	(—.10)
3	.469	.220	.44	.22
4	.409	.168	.18	.01
5	.677	.458	.55	.09
6	.895	.801	.77	(—.03)
7	.846	.716	.63	(—.09)
8	.658	.434	.64	.21
9	.844	.713	.71	.00

TABLE 3
Residuals, Disregarding Sign

Res.	Freq.
.00	6
.01	7
.02	13
.03	7
.04	1
.05	1
.06	
.07	1

Inspection of the correlations in Table 1 makes one suspect that the columns and rows are nearly proportional. The writer applied Spearman's formula for the explicit solution of single-factor loadings, which obviates the computation of tetrads. The centroid method with two successive adjustments in communalities gave the same single-factor solution for this problem as shown in Table 2. In the first column of Table 2 we have the nine tests by number. In the next column we have the first factor saturations, which account for the correlations. In Table 3 we have a frequency distribution of residuals, and it is doubtful whether any one would be tempted to extract a second factor from residuals so small as these—to say nothing of pulling out nine factors here.

In the third column of Table 2 we have the communalities. The reliabilities r_{jj} which are listed in the table were given by Davis. Surely we should all agree that the test reliabilities show the maximum part of the total variance that can be interpreted in terms of reading skills. The complement $(1 - r_{jj})$ is the variance in each test that is lost in the variable errors. The difference between the reliability and the communality is the specific variance which is unaccounted for. This difference $(r_{jj} - h_j^2)$ is listed in the table. It should be noted that six of these values are very small. The several small negative values are due, no doubt, to experimental variation. For these six tests there is no variance left to explain. For these six tests

the one common factor accounts for the entire available variance indicated by the reliability coefficients. The three remaining tests (1, 3, and 8) have specific variances of .25, .22, and .21. These represent fractions of test variances that are not shared with the other tests in the battery. Kelley objects to the use of communalities in factor analysis because they leave some unique variance for each test. But reliabilities that are less than unity demonstrate the existence of unique variance in each test.

If these were my data, I should write the following conclusions:

(1) The given correlations are accounted for by a single common factor with remarkably small residuals. This is an outstanding result in view of the fact that the tests were constructed to represent nine supposedly different skills.

(2) The nature of the tests indicates that the one common factor is reading ability, which is not surprising since all the tests were intended to appraise reading skills.

(3) Six of the tests show no further specific variance, since their communalities are practically equal to their reliabilities.

(4) Since these nine tests are shown to be measures of the same reading function, we have here no evidence about the components of the complex that we call reading ability. That question still remains to be investigated by new tests in the hope of identifying fundamental parameters of reading ability.

PSYCHOMETRIKA—VOL. 11, NO. 3
SEPTEMBER, 1946

BOOK REVIEWS

JOHN G. DARLEY, *Testing and Counseling in the High School Guidance Program.* Chicago: Science Research Associates. 1943. Pp. 222.

A practical volume of workable guidance suggestions comes from the personnel laboratories of the University of Minnesota, pioneer and leader in guidance in higher education The author, John G. Darley, director of the Testing Bureau, indicates the general character of the discussion by a statement in the preface: "This book is written primarily for teachers on the job who work with students and for school administrators who want to understand what to expect of counseling." The treatment throughout is nontechnical, simple, and functional. The style is lively and informal. It is possible, however, that the author may have underestimated the capabilities of teacher-counselors. A carefully worded simplified introduction to each topic might well have been followed by a more critical consideration of some of the significant issues and problems related to each area of discussion.

The book deals with only one phase of personnel work, individual counseling. This important function is given rather full consideration on a descriptive and practical level. A sizeable section is devoted to the use of objective tests in counseling. The basic statistical procedures are explained in an exceptionally clear, common-sense manner with a minimum stress on mathematical formulas and processes. For example, student achievement is diagnosed by noting the position of each individual on a correlation table for which the variables are scholastic ability and the mastery of curriculum content. The problem cases of the underachievers and over-achievers are located for study and counseling. The importance of sampling is rather well presented.

A praiseworthy feature of the discussion is the occasional comment on the shortcomings of measuring devices available to counselors. "Tests are tools which are harmful as well as useful in their application to the student, depending upon the person who uses or interprets the test." The relatively low predictive efficiency of most correlation coefficients is realistically presented. The problem of test validity is stated quite frankly. In spite of the book's more than usual honesty concerning the limitations of tests, this reviewer believes that a more complete treatment of test reliability is needed. Actual examples of student scores should have been furnished with accompanying sampling-error ranges. The fact that every measurement must be interpreted as a zone rather than a point should have been stressed. The high degree of validity required for individual prediction should have been contrasted with that sufficient for group selection or placement as in hiring workers for an industrial plant.

The selection of suitable tests is treated in a practical manner with a few specific examples briefly summarized. Here again an even greater realism in picturing the strengths and weaknesses of actual tests might well have been employed. Obviously such a viewpoint is debatable. Few counselors, however, possess a balanced and informed understanding of the real worth of existing measuring devices.

Many modern educators would, no doubt, question the emphasis on published standardized tests as opposed to teacher-made objective achievement tests. Any considerable dependence on published tests in content fields in high school (in contrast to specific skill areas) reflects a rather conservative curriculum policy, since such dependence tends to standardize both subject-matter and teaching approach.

A much needed word of caution concerning the interpretation of personality tests by teacher-counselors appears. This word could with justice have been elaborated into several paragraphs. Too many amateur psychologists show an unwarranted optimism in passing judgment on personality quirks and difficulties from group-type, paper-and-pencil personality measures.

The last third of the book contains a useful analysis of the most frequent types of student problems. The common-sense approach reflects much mature experience in counseling on the part of the author. A rather complete chapter outlines some of the elements important to the development of skill in interviewing. Another current controversy is suggested by this discussion. An influential group of clinical workers and counselors will accuse the author of presenting a counselor-centered rather than a client-centered approach to the student. The relative merits of different counseling techniques under varying circumstances and with different objectives in mind should be clearly explained to the beginning guidance worker. There is undoubtedly an appropriate place for an informational interview as well as for the so-called nondirective approach.

The final chapter encourages the administrators and teachers of smaller high schools in rural areas and less populous towns by describing the evolution of a successful guidance program in North Dakota. These concrete experiences appear to justify the author's enthusiasm and general optimism regarding a counseling program for the average high school in America.

WELTY LEFEVER
The University of Southern California